Handbook of Deontic Logic and Normative Systems

Volume 2

Handbook of Deontic Logic and Normative Systems

Volume 2

Edited by

Dov Gabbay
John Horty
Xavier Parent
Ron van der Meyden
Leendert van der Torre

ISBN 978-1-84890-363-0

College Publications
Scientific Director: Dov Gabbay
Managing Director: Jane Spurr

http://www.collegepublications.co.uk

Cover produced by Laraine Welch

Table of Contents

Preface 1
Dov Gabbay, John Horty, Xavier Parent, Ron van der Meyden and Leendert van der Torre

Further Background

Preference Semantics for Hansson-type Dyadic Deontic Logic: A Survey of Results 7
Xavier Parent

Recent Thought on *Is* and *Ought*: Connections, Confluences and Rediscoveries 71
Lloyd Humberstone

More Concepts and Problems

Logics for Supererogation and Allied Normative Concepts 155
Paul McNamara

Deontic Logic and Changing Preferences 307
Johan van Benthem and Fenrong Liu

More New Frameworks

Adaptive Deontic Logics 367
Frederik Van De Putte, Mathieu Beirlaen and Joke Meheus

External Relations

Practical Reasoning: Problems and Prospects 463
Richmond H. Thomason

Deontic Logic and Natural Language 499
Fabrizio Cariani

Deontic Logic and Ethics 549
Shyam Nair

Logic and the Law: Philosophical Foundations, Deontics, and Defeasible Reasoning 657
Guido Governatori, Antonino Rotolo and Giovanni Sartor

Deontic Logic and Game Theory 765
Olivier Roy

Preface

Eight years ago, we published the first volume of the *Handbook of Deontic Logic and Normative Systems*, with the goal of providing an overview of the main lines of research on deontic logic and related topics. We now publish a second volume. While not neglecting historical work, this volume, like the first, concentrates on the significant advances in deontic logic that have occurred during the past three decades, or roughly since 1990. These changes have resulted largely, though not entirely, from the interaction of deontic logic with a variety of other fields outside of its traditional home in philosophy, including computer science, legal theory, linguistics, and economics.

As editors, we have been guided by four ideas, already articulated in our introduction to the first volume, but repeated here. First, we have tried to highlight new developments, and new prospects for deontic logic. Second, we have tried to combat the impression that deontic logic exists only as a collection of abstract formal systems, sometimes lacking in motivation. Instead, we want to emphasize the real problems that give rise to the formalisms developed by deontic logicians, as well the potential for real applications in a variety of fields. Third, we have made every effort to provide authors with the freedom to present their material in depth, sometimes resulting in chapters of monographic length and scope, containing the first comprehensive treatments of their subjects. Finally, we wanted the work to be affordable for individual researchers, not simply for those institutions willing to pay the exorbitant prices charged by commercial publishers, and even by certain commercial ventures masking as university presses. For this reason, we chose to work with College Publications, a non-profit publisher run by academics and for academics. We recommend this service to others.

Although the Handbook was, in a certain narrow sense, managed by a group of editors, it is, more accurately, the work of community. The community is the Society of Deontic Logic and Normative Systems, which sponsors the International Conferences on Deontic Logic and Normative Systems; these conferences, generally known as DEON meetings, occur biennially (except when the regular biennial meeting schedule happens to be interrupted by a pandemic). The current volume was the subject of extensive discussion at the DEON meetings in Ghent in 2014, in Bayreuth in 2016, and in Utrecht in 2018. Many chapters were first conceived of at these meetings, and appropriate authors identified (and

Handbook of Deontic Logic and Normative Systems, Volume 2

in some cases, pressured). Then, a process began: drafts were sent to external readers, discussed among editors, revised, sometimes reviewed again, until finally ready for publication.

We are grateful to the external readers who reviewed manuscripts for this volume: Guillaume Aucher (twice), Hein Duijf, Lou Goble (twice), Davide Grossi, Sven Ove Hansson, Reka Markovich, Henry Prakken, Mark Schroeder, Allard Tamminga, and Malte Willer.

We are also especially grateful to Jane Spurr of College Publications, a model of patience, competence, and kindness. Without Jane's oversight, this volume would not exist, and neither would the first volume, and neither would many of the other books that have meant so much to the logical community over the past thirty years.

As the editors met in 2008 and 2009 to discuss the shape of these volumes, we had an elegant overall architecture in mind. The chapters from the first volume were to be divided into a first part, Background, covering historical aspects of deontic logic, a second part, Concepts and Problems, covering substantive issues in deontic logic, and a third part, New Frameworks, devoted to new theoretical approaches in the field; the second volume, it was thought, would be devoted entirely to a number of applications, linking deontic logic with other fields. Having made this elegant plan, we then learned, once again, the futility of making plans. For one reason or another, several promised chapters could not be completed in time for the first volume and had to be included in the second; other promised chapters (including chapters promised by the editors) have not yet been completed at all. More happily, new chapters that we had not originally envisaged came to light, including one that appeared fully-formed and very nearly out of the blue.

As a result, we had to abandon our elegant overall architecture. The current volume begins with several oddly-titled parts that simply echo the main chapter divisions of the previous volume. The first part, Further Background, includes a chapter on preference semantics for dyadic deontic logic and a chapter on the is/ought problem, going back to Hume, of course, but subject to sustained and increasingly formal discussion over the past six decades. The second part, More Concepts and Problems, includes a chapter on supererogation and related normative concepts and then a chapter investigating logics governing agents with changing preferences. The third part, More New Frameworks, contains a chapter developing deontic logic within the framework of adaptive logic.

These initial parts of the current volume contain material that would more naturally have appeared in the first volume. It is only with the final

part, External Relations, that we turn at last to the applications that were supposed to be the central focus of the current volume. This part contains chapters on deontic logic as it relates to practical reasoning, to natural language, to ethics, to legal reasoning, and to game theory.

As we slowly assembled the current volume, the subject of deontic logic itself has not stood still, and we now see the need for other chapters exploring still other applications—for example, deontic logic in the formalization of the rights relation, epistemic permissions and obligations, reactive deontic logics, deontic logic and decision theory, implementations of deontic logics in theorem provers, deontic description logics, applications of deontic logic in planning, deontic logic in argumentation theory, and the role of deontic logics in machine ethics. At the same time as we imagine these new chapters focusing on recent developments, we are also aware that several topics essential to any adequate coverage of the field have yet to find appropriate authors—these include contrary-to-duty obligations, the interaction of deontic logics with logics of time and action, and connections between deontic logics formulated in the standard possible-worlds framework and in the imperative framework.

We therefore anticipate that there will have to be a third volume to this series. Fortunately for us, and perhaps fortunately for the field, we also anticipate that this third volume will be managed by a new set of editors.

Dov Gabbay
John Horty
Xavier Parent
Ron van der Meyden
Leon van der Torre

Note

Chapters 2, 4, 5 and 6 included here were previously published in the *Journal of Applied Logics — The IfCoLog Journal of Logics and their Applications*, and are reprinted here by agreement with the Journal.

Further Background

1

Preference Semantics for Hansson-type Dyadic Deontic Logic: A Survey of Results

Xavier Parent

Abstract. This chapter discusses the Hansson-type preference semantics for dyadic deontic logics. In that framework the conditional obligation operator is interpreted in terms of best antecedent-worlds. I survey results pertaining to the meta-theory of such logics, focusing on axiomatization issues. The goal is to provide a "roadmap" of the different systems that can be obtained, depending on the special properties envisaged for the betterness relation, and depending on how the notion of "best" is understood (optimality *vs.* maximality, stringent *vs.* liberal maximization). In addition, the systems' decidability and automated theorem-proving for them are discussed, and variant truth-conditions for the conditional obligation operator are reviewed.

1 Introduction **8**

2 Syntax and semantics **13**
2.1 Syntax 13
2.2 Semantics–basic setting 14
2.3 Two notions of "best" 15
2.4 Properties of $\succeq$ 17
2.5 Where the opt rule *vs.* the max rule makes a difference . 19
2.6 Selection functions 21

This chapter is dedicated to the memory of Lennart Åqvist, who died on 7 March 2019 at the age of 86. This work was supported by WWTF MA16-028. I am indebted to Lou Goble for his careful reading of the chapter and for valuable comments. I wish to thank Christoph Benzmueller and Alexander Steen for feedback on the section devoted to automated theorem-proving, and Jose Carmo and Dov Gabbay for feedback on the section dealing with decidability. I also would like to extend my gratitude to Cleo Condoravdi for helpful discussions on Kratzer, and to Walter Bossert for useful discussions on rational choice theory and on Theorem 4.15. Last, I would like to thank those who have commented on this essay, or aspects of it, at various stages of its development. In particular I would like to mention Richard Booth, Joerg Hansen, Sven Ove Hansson, Jeff Horty, Paul McNamara, David Makinson and Leon van der Torre.

Handbook of Deontic Logic and Normative Systems, Volume 2

3 Proof systems **24**
3.1 Mixed alethic-deontic logics 24
3.2 Pure deontic conditional logics 27

4 Determination results **30**
4.1 Core results 30
4.2 Adding reflexivity and totalness 31
4.3 Transitivity without smoothness (max rule) 33
4.4 Pure deontic conditional counterparts 34
4.5 Methods for proving completeness 35
4.5.1 Direct canonical model construction 36
4.5.2 Completeness-via-selection-functions 37

5 Decidability and automated theorem-proving **41**
5.1 Decidability 41
5.2 Automated theorem proving 45

6 Alternative truth-conditions **48**
6.1 The Danielsson-van Fraassen-Lewis truth-conditions . . . 49
6.2 The Burgess-Boutilier-Lamarre truth-conditions 52

7 Conclusion **56**

Appendix A: Proof of Thm 3.3 (vi) **63**

Appendix B: Proof of Thms 4.2 (ii) and 4.5 (ii) **64**

Appendix C: Proof of Thms 4.7 and 4.8 **67**

1 Introduction

Beginning with work by Danielsson [1968] and Hansson [1969], so-called Dyadic Deontic Logic (hereinafter referred to as "**DDL**") aims at providing a formal analysis of conditional obligation sentences within a preference-based semantics. The language of **DDL** employs a dyadic (or conditional) obligation operator $\bigcirc(-/-)$, where $\bigcirc(B/A)$ is read as "It is obligatory that B, given that A". This construct is interpreted using a preference relation, which orders all the possible worlds in terms of comparative goodness or betterness. In that framework $\bigcirc(B/A)$ is taken to hold, whenever all the best A-worlds are B-worlds.

DDL is a natural generalization of Monadic Deontic Logic (hereinafter referred to as "**MDL**"). The semantics of this one uses a binary classification of possible worlds into good/bad. For **DDL**, this binary classification is relaxed to allow for grades of ideality between these two extremes.[1] This leads to the use of a conditional obligation operator that is primitive rather than being defined in terms of the standard (monadic) obligation operator and some other familiar constructs like material implication or strict implication.

DDL uses the possible world semantics in novel ways with a view to solving issues related to two different kinds of deontic conditionals:

Contrary-to-duty conditionals Since the publication of Chisholm [1963], deontic logicians have struggled with what has become known as the "contrary-to-duty" (CTD) problem. It is the problem of giving a formal treatment to those obligations—called "contrary-to-duty" by Chisholm—which come into force when some other obligation is violated. **DDL** was initially developed in order to handle this first type of deontic conditional. According to Hansson and others, like van Fraassen [1972] and Lewis [1973; 1974], the problems raised by CTDs call for the use of an ordering on possible worlds, in terms of preference or relative goodness, and **MDL** fails in as much as its semantics does not allow for grades of ideality.

Defeasible deontic conditionals Independently of the above, the use of a preference relation has also been advocated in relation to the analysis of the notion of defeasible conditional obligation. In particular, Alchourrón [1993] argues that preferential models provide a better treatment of this notion than the usual Kripke-style models do. Indeed, a defeasible conditional obligation is one that leaves room for exceptions. Under a preference-based approach, we no longer have the deontic analogue of two laws, the failure of which constitutes the main formal feature expected of defeasible conditionals. One is "deontic" modus-ponens, also known as Factual Detachment (FD): $\bigcirc(B/A)$ and A imply $\bigcirc B$. The other is

[1] A remark on my choice of name is in order. **MDL** is more commonly known as "Standard Deontic Logic" (SDL), and **DDL** as "Dyadic Standard Deontic Logic" (DSDL). Both names appear in Hansson's seminal paper. Throughout this chapter I will not use the label SDL, because it tends to carry the connotation that the framework in question is still a recognized "standard". As Hilpinen and McNamara [2013, p. 38] point out, to call SDL a standard is a misnomer. **MDL** refers to a family of systems, which were called **D**, **DS4**, **DM** and **DS5** by Hanson [1965]. (Other labels have been used in the literature.)

the law of Strengthening of the Antecedent (SA): $\bigcirc(B/A)$ entails $\bigcirc(B/A \wedge C)$.

There is an extensive literature on the treatment of these notions within a preference-based framework. Regarding contrary-to-duties, the reader may wish to consult [van Fraassen, 1972; Lewis, 1973; Tomberlin, 1981; Loewer and Belzer, 1983; Kratzer, 1991; Prakken and Sergot, 1997; van der Torre and Tan, 1999; Hilpinen and McNamara, 2013]. Concerning defeasible conditional obligations, the reader is referred to [Makinson, 1993; Boutilier, 1994; Alchourrón, 1995; Asher and Bonevac, 1997; van der Torre and Tan, 1997; Horty, 2014]. It is not the purpose of this chapter to evaluate such treatments, nor is it to discuss the relationship between dyadic deontic logic and frameworks developed in other closely related areas, like revealed preference theory (as introduced by the economist Samuelson), the logic of conditionals (as developed in the 1970's following Stalnaker and Lewis), or the theories of nonmonotonic inference operations (as constructed in the 1980's in the context of logics for artificial intelligence). All these frameworks share the idea of using a semantics based on a notion of minimality under a preference relation, or equivalently, a notion of maximality under its converse. For a good discussion of the interplay between these areas, the reader is referred to [Makinson, 1993].[2]

The aim of this chapter is to present a survey of recent results pertaining to the meta-theory of **DDL**. Since the publication of Hansson's seminal paper, substantial contributions have been made to enhance our understanding of the meta-theory of **DDL**, starting with work by Spohn [1975], and continuing with work by Åqvist [1987; 1993; 2002], Hansen [1999], Goble [2015; 2019] and myself [Parent, 2008; Parent, 2010; Parent, 2014; Parent, 2015]. However, there is still no systematic survey of the field. The present chapter aims at filling in this gap. The goal is to provide a "roadmap" of the different systems that can be obtained, based on two types of considerations or variations.

The first type of consideration is familiar from modal logic. Different systems can be obtained by varying the conditions on the preference relation. In general the imposition of a condition has the effect of validating a modal formula. In monadic modal logic, we have a clear picture of the different systems that can be obtained depending on the properties of the accessibility relation. In dyadic deontic logic, this picture is still missing. Results in the literature have so far mostly concerned classes of

[2]Makinson does not discuss the connection with rational choice theory. This one is examined by [Rott, 2001] among others.

structures with strong conditions on the betterness relation. One such condition is the property of transitivity, which has been called into question by moral philosophers and economists.[3] One would like to know what happens when such a condition is relaxed. What Lewis [1973] calls the limit assumption is another requirement that one would like to be able to drop. Roughly speaking, it says that a set of possible worlds should always have a best element. A number of deontic logicians objected to the limit assumption, Lewis [1973, p. 97-98] among them. It is not widely known what happens when these properties are not assumed.

This brings into the forefront so-called correspondence theory, devoted to the systematic study of relations between classes of frames and modal languages. Van Benthem [2001, §3.2] asks if, or to what extent, such a theory can be developed for conditional logic. Such a study falls outside the scope of the present chapter. But I hope the considerations it offers can be used as a stepping stone towards the development of such a theory.

The second type of consideration this chapter introduces concerns the notion of "best", in terms of which the truth conditions for $\bigcirc(-/-)$ are typically phrased. One can distinguish between two ways to understand the notion of a world being best: it can be either optimal or maximal. This distinction is well-known from rational choice theory where most authors follow Herzberger [1973] in using the terms "stringent" *vs.* "liberal" maximization for what (following Sen [1997]) I call optimality *vs.* maximality. For some item x to qualify as an optimal element of X, it must be at least as good as every member of X. For x to count as a maximal element, no other element in X must be strictly better than it. Thus, while the optimal elements are all equally good, the maximal elements are either equally good or incomparable. Depending on what notion of "best" is used, one gets different truth conditions for $\bigcirc(-/-)$, but also different forms of the limit assumption.

I remark in passing that there is some variation in terminology. For instance, [Bossert and Suzumura, 2010] prefer the labels "maximal *vs.* greatest" element rationalizability. On the other hand, the choice to use "optimal" and "maximal" the way just described is not mine, but Sen's (see in particular [Sen, 1997, §5]). I have heard people swap the two terms, and take optimal as meaning "not-dominated", and maximal as meaning "dominates-all-others". (See, *e.g.*, the definition of optimal in [Horty, 2001, p. 72].) In the end, it does not matter which way we speak, so long as we understand and agree on what we mean and do not

[3] Cf. [Sen, 1971] and [Temkin, 1987].

allow the coexistence of two conflicting ways of speaking to engender confusion. In this chapter I will stick to Sen's terminology.

This investigation takes place in the conditional logic setting put forth by Åqvist in a series of publications [Åqvist, 1987; Åqvist, 1993; Åqvist, 2002] rather than in Hansson's original setting. That one is studied axiomatically by Spohn [1975] and Goble [2019]. Readers should be warned that there is far less standardization in preference semantics than in the usual Kripke-style semantics for deontic logic, and more room for variation. This is due to the fact that there are several factors that must be juggled all at once. Thus, even when sticking with Åqvist's approach, more semantical variations than the above two can be made. For instance, under the Åqvist account the ranking is not world-relative. However, as Makinson [1993] points out, one may want to allow for the ranking to vary across possible worlds. This extra choice (and some others) are studied axiomatically by Goble, who in his [2015] pursues a similar project. It falls outside the scope of the present paper to integrate his results. The present chapter is not, and does not pretend to be, a comprehensive survey of Hanssonian approaches to dyadic deontic logic, so much as a summary of certain results, mainly my own, that would help the reader understand some important aspects of the Hanssonian approach, but does not address the scope of that approach from either a formal or philosophical point of view.

As part of motivating the formal moves to be developed next, I briefly recall how the framework handles the standard CTD scenarios, like Chisholm's paradox.

Example 1.1. *[Chisholm's scenario] Consider the following set of sentences, where h can be read as the fact that a certain man goes to the assistance of his neighbors and t as the fact that he is telling them that he is coming:*

$$\Gamma = \{\bigcirc h, \bigcirc(t/h), \bigcirc(\neg t/\neg h), \neg h\}$$

$\bigcirc h$ expresses what is usually called a primary obligation. $\bigcirc(\neg t/\neg h)$ is its associated CTD obligation, and $\bigcirc(t/h)$ is its associated ATD (according-to-duty) obligation. Figure 1 describes a typical preference model of Γ. Here the convention is that at each world $a \in W$, I list the propositional letters that a satisfies, omitting those that it makes false. The best overall world is the one where both h and t hold, and the worst overall world is the one where t holds but h does not. In between one sees two worlds, one with h but not t and the other with neither h nor t. All the formulas in Γ are satisfied in a_3 and a_4. This shows that the set Γ is consistent. The primary obligation holds, because the best overall

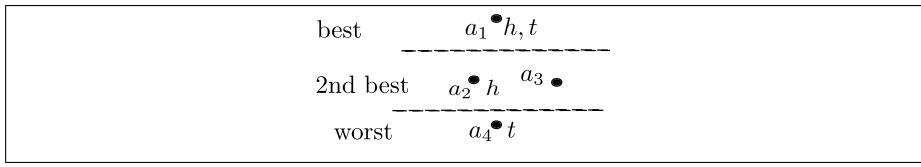


Figure 1: A typical model of Chisholm's scenario

word satisfies h. The CTD obligation holds, because the best $\neg h$-world satisfies $\neg t$. The ATD obligation holds, because the best h-world satisfies t. It is worth mentioning that this approach to the CTD scenarios only works because neither (FD) nor (SA) are valid under this approach, as the model of Figure 1 demonstrates.[4]

The layout of this chapter is as follows. In Section 2, the syntax and the semantics are described. In Section 3, the relevant proof systems are introduced. In Section 4, the determination results available at the time of writing this chapter are reviewed. In Section 5, the decidability of the theoremhood problem is established, and automated theorem-proving is discussed. In Section 6 variant truth-conditions are reviewed. Section 7 concludes. Supplementary material is gathered in three appendices. In particular, the proof of two new results is given.

2 Syntax and semantics

2.1 Syntax

Definition 2.1. *The language $\mathcal{L}$, or set of well-formed formulas (wffs), is generated from a set $\mathbb{P}$ of propositional atoms by the following BNF:*

$$A ::= p \in \mathbb{P} \mid \neg A \mid A \vee A \mid \Box A \mid \bigcirc(A/A)$$

$\neg A$ is read as "not-A", and $A \vee B$ as "A or B". $\Box A$ is read as "A is settled as true", and $\bigcirc(B/A)$ as "B is obligatory, given A". A is called the antecedent, and B the consequent.

The following derived connectives are introduced. $P(B/A)$ ("B is permitted, given A") is short for $\neg\bigcirc(\neg B/A)$, $\bigcirc A$ ("A is unconditionally obligatory") and PA ("A is unconditionally permitted") are short for $\bigcirc(A/\top)$ and $P(A/\top)$, respectively. $\Diamond A$ is short for $\neg\Box\neg A$. Other Boolean connectives are defined as usual.

[4](FD) yields $\bigcirc\neg t$. This "contradicts" the fact that the best overall world satisfies t, so that $\bigcirc t$ holds. (SA) warrants the move from $\bigcirc t$ to $\bigcirc(t/\neg h)$. This "contradicts" the third formula in Γ.

Åqvist's language goes beyond Hansson's by including alethic modalities, mixed formulas (in which deontic formulas are combined with Boolean ones) and iterated deontic modalities.

2.2 Semantics–basic setting

Definition 2.2 (Preference model)**.** *A preference model is a structure*

$$M = (W, \succeq, v)$$

in which

(i) $W \neq \emptyset$ *(W is a non-empty set of "possible worlds");*

(ii) $\succeq \subseteq W \times W$ *(intuitively,* $\succeq$ *is a betterness or comparative goodness relation; "*$a \succeq b$*" can be read as "world a is at least as good as world b");*

(iii) $v : \mathbb{P} \to \mathcal{P}(W)$ *(v is an assignment, which associates a set of possible worlds to each propositional atom p).*

$\succ$ denotes the strict relation induced by $\succeq$, defined as its "strengthened converse complement" and obtained by putting $a \succ b$ whenever $a \succeq b$ and $b \not\succeq a$. $a \succ b$ may be read as "a is strictly better than b". Note that $\succ$ is by definition irreflexive (*i.e.*, for all a, $a \not\succ a$). Two worlds a and b are said to be equally good or indifferent, $a \equiv b$, whenever $a \succeq b$ and $b \succeq a$. They are said to be incomparable, $a||b$, whenever $a \not\succeq b$ and $b \not\succeq a$.[5]

Definition 2.3 (Satisfaction relation)**.** *Given a model* $M = (W, \succeq, v)$ *and a world* $a \in W$*, the satisfaction relation* $M, a \vDash A$ *(read as "world a satisfies A in model M") is defined by induction on the structure of A. The clauses are as usual for the Boolean connectives and* $\Box$*:*

$$M, a \vDash p \text{ iff (if and only if) } a \in v(p)$$
$$M, a \vDash \neg A \text{ iff } M, a \nvDash A$$
$$M, a \vDash A \vee B \text{ iff } M, a \vDash A \text{ or } M, a \vDash B$$
$$M, a \vDash \Box A \text{ iff } \forall b \;\; M, b \vDash A$$

The clause for the dyadic obligation operator is:

$$M, a \vDash \bigcirc(B/A) \quad \text{iff} \quad \text{best}_{\succeq}(\|A\|^M) \subseteq \|B\|^M$$

[5] The betterness relation $\succeq$ may be defined in terms of some more basic ingredients in the semantics. (See, for instance, [Kratzer, 2012] and [Prakken and Sergot, 1997]). However, most articles in the field do not consider this course, and neither will I in this chapter. Kratzer's theory is discussed in more detail in the chapter in this volume "Deontic logic and natural language" by F. Cariani.

As usual $\|A\|^M$ denotes the truth-set of A (*i.e.*, the set of worlds at which A holds). The notation $\text{best}_{\succeq}(\|A\|^M)$ is a shorthand for the set of best (according to $\succeq$) worlds in which A is true. Intuitively, $\bigcirc(B/A)$ holds at a whenever all the best A-worlds are B-worlds. Note that, by definition of $P(-/-)$, $M, a \vDash P(B/A)$ iff $\text{best}_{\succeq}(\|A\|^M) \cap \|B\|^M \neq \emptyset$. Intuitively: $P(B/A)$ holds whenever at least one best A-world is a B-world. I will postpone the definition of $\text{best}_{\succeq}(\|A\|^M)$ until the next section. When the context allows, I will drop the symbol M and just write $\|A\|$ and $a \models A$.

The notions of semantic consequence, validity and satisfiability are defined as usual.

2.3 Two notions of "best"

As mentioned in Section 1, there are two ways to formalize the notion of best antecedent-worlds: one may do it using the notion of optimality, or the notion of maximality.[6] They are not clearly distinguished in the deontic logic literature even though their differences can be significant. They may be defined thus:

$$\text{opt}_{\succeq}(\|A\|^M) = \{b \in \|A\|^M \mid \forall c \ \ (c \vDash A \rightarrow b \succeq c)\}$$
$$\max_{\succeq}(\|A\|^M) = \{b \in \|A\|^M \mid \forall c \ \ ((c \vDash A \ \& \ c \succeq b) \rightarrow b \succeq c)\}$$

Maximality can equivalently be defined in terms of $\succ$:

$$\max_{\succ}(\|A\|^M) = \{b \in \|A\|^M \mid \nexists c \ \ (c \vDash A \ \& \ c \succ b)\}$$

It is easy to see that $\text{opt}_{\succeq}(\|A\|^M) \subseteq \max_{\succeq}(\|A\|^M)$ although the converse inclusion may fail. Typically, it will fail if there are "gaps" in the ranking.

Example 2.4 (Gaps)**.** *Define $M = (W, \succeq, v)$, with $W = \{a, b\}$, $v(p) = W$, and $\succeq = \{(a, a), (b, b)\}$. We have $a \| b$.* $\max_{\succeq}(\|p\|^M) = \{a, b\}$ *but* $\text{opt}_{\succeq}(\|p\|^M) = \emptyset$.

Totalness of $\succeq$ ("for all $a, b \in W, a \succeq b$ or $b \succeq a$") may be shown to be a sufficient condition for the two notions of "best" to coincide. We have already seen that $\text{opt}_{\succeq}(\|A\|^M) \subseteq \max_{\succeq}(\|A\|^M)$. Now,

Observation 2.5. $\max_{\succeq}(\|A\|^M) = \text{opt}_{\succeq}(\|A\|^M)$ *if $\succeq$ is total.*

Proof. The right-in-left inclusion holds by definition. The left-in-right inclusion calls upon totalness. To see why, assume $\succeq$ is total, and let

[6] As mentioned, I adopt this terminology from Sen [1997].

$a \in \max_{\succeq}(\|A\|^M)$. Consider $b \in \|A\|^M$. By totalness, $a \succeq b$ or $b \succeq a$. In the second case, $a \succeq b$, since $a \in \max_{\succeq}(\|A\|^M)$. Either way, $a \succeq b$, and so $a \in \mathrm{opt}_{\succeq}(\|A\|^M)$. □

Thus, one gets two different pairs of evaluation rules depending on which of the following two equations is adopted:

$$\mathrm{best}_{\succeq}(\|A\|^M) = \max_{\succeq}(\|A\|^M) \qquad \text{(max rule)}$$

$$\mathrm{best}_{\succeq}(\|A\|^M) = \mathrm{opt}_{\succeq}(\|A\|^M) \qquad \text{(opt rule)}$$

Both definitions can be found in the literature.[7] From now onward, I will refer to the first equation (*resp.* second equation) as the max rule (*resp.* opt rule). From Observation 2.6, it immediately follows that, in a given model M with $\succeq$ total, the same deontic formulas are true at a given world whatever rule is used.

This chapter focuses on the above two definitions of "best". As a matter of fact, variant definitions have been proposed. The purpose of these variations is often to remedy the emptiness of the set of best worlds when the betterness relation admits cycles, like in Figure 2. Condorcet's well-known voting paradox [Sen, 1969] is often used to show the plausibility of this kind of situations.

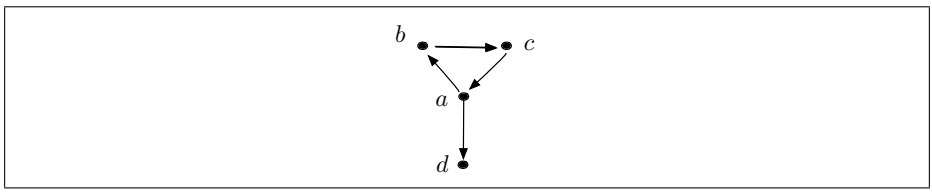


Figure 2: A top cycle. An arrow from a to b represents $a \geq b$. No arrow from b to a means $b \not\geq a$.

Hansson [2009] suggests maximizing with respect to the transitive closure $\succeq^\star$ rather than $\succeq$ itself.[8] Recall that $a \succeq^\star b$ iff there are $c_1, ..., c_n$

[7] For instance, Hansson [1969], Makinson [1993, §7.1], Schlechta [1995], Prakken and Sergot [1997], van der Torre and Tan [1997, p. 95], Horty [2001, p. 72] and Stolpe [2020] use the max rule. In contrast, Spohn [1975], Åqvist [1987; 2002], Fehige [1994, p. 43], Alchourrón [1995, p. 76], McNamara [1995], Hansen [2005, §6] work with the opt rule. Neither Goldman [1977], nor Jackson [1985], nor Hilpinen [2001, §8.5] specifies what notion of "best" is meant. (The last one uses "best" and "deontically optimal" interchangeably, but leaves optimality undefined.)

[8] There is room for variation here. Hansson considers four alternative constructions, and finally settles on that one.

such that $a \succeq c_1 \succeq ... \succeq c_n \succeq b$. I will call this variant the "quasi-maximality" (quasi-max, for short) rule:

$$\text{best}_{\succeq}(\|A\|^M) = \max_{\succeq^\star}(\|A\|^M) \qquad \text{(quasi-max rule)}$$

where

$$\max_{\succeq^\star}(\|A\|^M) = \{b \in \|A\|^M \mid \forall c \ ((c \vDash A \ \& \ c \succeq^\star b) \rightarrow b \succeq^\star c)\}$$

It is worth noticing that, if $\succeq$ is transitive, then $\succeq = \succeq^\star$, so that the quasi-max rule coincides with the original max rule:

Observation 2.6. $\max_{\succeq}(\|A\|^M) = \max_{\succeq^\star}(\|A\|^M)$ *if* $\succeq$ *is transitive.*

A thorough study of such alternative definitions must be postponed to another occasion. I will report a completeness result for the interpretation under the quasi-max rule in Section 4.3.

2.4 Properties of $\succeq$

The properties usually envisaged for $\succeq$ are reflexivity, transitivity, totalness, and the so-called limit assumption. The first three may be given the form:

- reflexivity: for all $a \in W, a \succeq a$;
- transitivity: for all $a, b, c \in W$, if $a \succeq b$ and $b \succeq c$, then $a \succeq c$;
- totalness: for all $a, b \in W, a \succeq b$ or $b \succeq a$.

The exact formulation of the limit assumption varies among authors. It can be given two basic forms:

Limitedness
If $\|A\| \neq \emptyset$ then $\text{best}_{\succeq}(\|A\|) \neq \emptyset$
Smoothness (or stopperedness)
If $a \vDash A$, then: either $a \in \text{best}_{\succeq}(\|A\|)$ or
$\exists b$ s.t. $b \succ a \ \& \ b \in \text{best}_{\succeq}(\|A\|)$

The name "limitedness" is from Åqvist [1987; 2002], "smoothness" from Kraus & al. [1990], and "stopperedness" from Makinson [1989]. Each of limitedness and smoothness may be specified further by identifying $\text{best}_{\succeq}(X)$ with either $\max_{\succeq}(X)$ or $\text{opt}_{\succeq}(X)$. A betterness relation $\succeq$ will be called "opt-limited" or "max-limited" depending on whether limitedness holds with respect to $\text{opt}_{\succeq}$ or $\max_{\succeq}$. Similarly, it will be called

"opt-smooth" or "max-smooth" depending on whether smoothness holds with respect to $\text{opt}_{\succeq}$ or $\max_{\succeq}$.[9]

This gives us four versions of the limit assumption. With the strong assumptions of transitivity and totalness, these different forms of the limit assumption coincide. However, with weaker constraints on $\succeq$, they may well diverge.

Theorem 2.7.

(a) (i) opt-limitedness implies max-limitedness;
(ii) given totalness of $\succeq$, max-limitedness implies opt-limitedness;
(b) (i) opt-smoothness implies max-smoothness;
(ii) given totalness of $\succeq$, max-smoothness implies opt-smoothness.

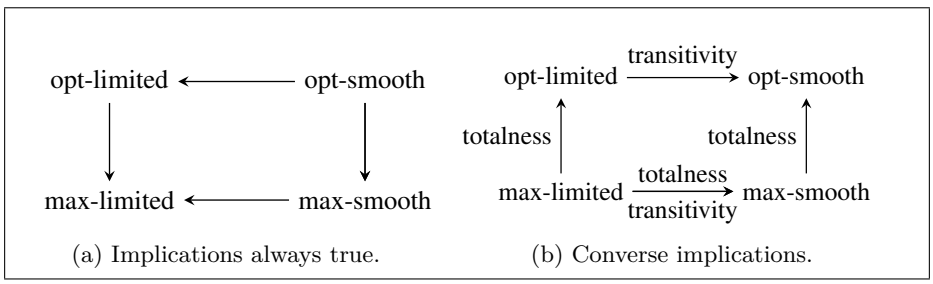
Proof. This follows at once from the definitions involved and Observation 2.6. □

Theorem 2.8.

(a) (i) max-smoothness implies max-limitedness;
(ii) given transitivity and totalness of $\succeq$, max-limitedness implies max-smoothness;
(b) (i) opt-smoothness implies opt-limitedness;
(ii) given transitivity of $\succeq$, opt-limitedness implies opt-smoothness.

Proof. See [Parent, 2014, Proposition 2]. □

Figure 3 represents the relationships just established in an Implication Diagram with the direction of the arrow representing that of implication. The implication relations shown in the picture on the left-hand



(a) Implications always true. (b) Converse implications.

Figure 3: Forms of the limit assumption, and their relationships.

[9]Hansson [1969] and Prakken and Sergot [1997] use max-limitedness, while Lewis [1974, p. 6], Spohn [1975], Åqvist [1987; 2002], Fehige [1994, p. 44], Alchourrón [1995, p. 84], McNamara [1995] and Hansen [2005, §6] use opt-limitedness, and Makinson [1993] and Schlechta [1995] max-smoothness. I am not aware of any authors who have considered opt-smoothness explicitly.

side hold without restriction. By contrast, those shown on the right-hand side hold under the hypothesis that $\succeq$ meets the property (or pair of properties) displayed as labeled.

In this chapter I only want to understand how the choice of a given version of the limit assumption affects the logic. The philosophical aspects of the limit assumption will not be discussed here−the reader should consult [Lewis, 1973; Fehige, 1994; McNamara, 1995; Hilpinen and McNamara, 2013]. Note that in linguistics the limit assumption has been given even more variant forms. (See, *e.g.*, the discussion in [Kaufmann, 2017].)

2.5 Where the opt rule *vs.* the max rule makes a difference

In this section, I give two examples of a valid formula for which the choice between the opt rule and the max rule makes a difference.

First, there is the example of the principle of rational monotony [Lehmann and Magidor, 1992], also called CV by Lewis [1973]. This is the principle

$$(P(B/A) \wedge \bigcirc(C/A)) \rightarrow \bigcirc(C/A \wedge B) \qquad \text{(RM)}$$

(RM) expresses a restricted principle of strengthening of the antecedent: one can strengthen an antecedent when the added condition B is permitted under the main condition A. Hence, doing the permitted has no effect on our other obligations.

Observation 2.9. *Under the opt rule, (RM) is valid if $\succeq$ is required to be transitive. Under the max rule, (RM) is valid if $\succeq$ is required to be both transitive and total.*

Proof. Assume that (i) $\mathrm{opt}_{\succeq}(\|A\|) \subseteq \|C\|$, (ii) $\mathrm{opt}_{\succeq}(\|A\|) \cap \|B\| \neq \emptyset$, and (iii) $\mathrm{opt}_{\succeq}(\|A \wedge B\|) \not\subseteq \|C\|$. From (iii), there is some a such that $a \in \mathrm{opt}_{\succeq}(\|A \wedge B\|)$ and $a \not\models C$. From (i), $a \notin \mathrm{opt}_{\succeq}(\|A\|)$, because $a \not\models C$. But $a \models A$. So there is some $b \models A$ with $a \not\succeq b$. From (ii), there is also some c such that $c \in \mathrm{opt}_{\succeq}(\|A\|)$ and $c \models B$. Since $c \models A \wedge B$, $a \succeq c$. Also, $c \succeq b$, since $c \in \mathrm{opt}_{\succeq}(\|A\|)$. By transitivity, $a \succeq b$. Contradiction. Hence, under the opt rule, (RM) is valid if $\succeq$ is transitive.

For the max rule, it suffices to invoke the above along with Observation 2.6. □

While under the opt rule transitivity is sufficient for the validity of law (RM), by contrast under the max rule it is not sufficient.

Observation 2.10. *There is a preference model $M = (W, \succeq, v)$, in which $\succeq$ is transitive, such that (RM) fails in M under the max rule.*

Proof. Put $M = (W, \succeq, v)$, with $W = \{a, b, c\}$, $\succeq = \{(a, b)\}$ and $v(p) = W$, $v(q) = \{b, c\}$ and $v(r) = \{a, c\}$. The model is depicted in Figure 4, where $\succeq$ is (vacuously) transitive. We have $\max_{\succeq}(\|p\|) = \{a, c\}$, $\max_{\succeq}(\|p \wedge q\|) = \{b, c\}$, $\|q\| = \{b, c\}$ and $\|r\| = \{a, c\}$. Under the max rule, (RM) fails, since $\bigcirc(r/p)$ and $P(q/p)$ hold while $\bigcirc(r/p \wedge q)$ does not (witness: b). □

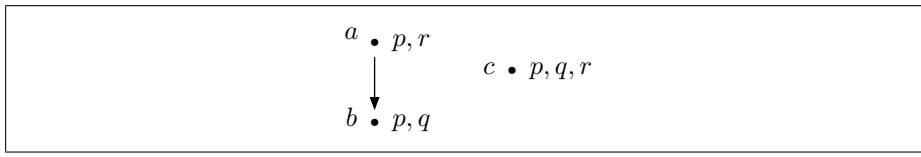


Figure 4: A countermodel to (RM)

What I say here about (RM) applies analogously to the following formula, named after Spohn [1975], who used it in his axiomatization of Hansson's system DSDL3:

$$(P(B/A) \wedge \bigcirc(B \to C/A)) \to \bigcirc(C/A \wedge B) \qquad \text{(Sp)}$$

We will see that (Sp) and (RM) are equivalent.[10] Spohn [1975, p. 247] himself argues that the assumption of totalness is iddle. He can do so only because he uses the opt rule instead of the max rule.

Here is my second example of a validity for which the choice between the opt rule and the max rule makes a difference:

$$P(A/A \vee B) \wedge P(B/B \vee C) \to P(A/A \vee C) \qquad (\gg\text{-trans})$$

($\gg$-trans) expresses a principle of transitivity for a notion of weak preference over formulas given by $A \gg B =_{def} P(A/A \vee B)$.[11] This says that A is ranked as at least as high as B iff it is permitted that A on the condition that either A or B.

Observation 2.11. *Under the opt rule, ($\gg$-trans) is valid if $\succeq$ is required to be transitive. Under the max rule, ($\gg$-trans) is valid if $\succeq$ is required to be both transitive and total.*

[10] Cf. Theorem 3.3 in Section 3.1.
[11] Cf. [Lewis, 1973, p. 54].

Proof. Assume that (i) $\text{opt}_{\succeq}(\|A \vee B\|) \cap \|A\| \neq \emptyset$, (ii) $\text{opt}_{\succeq}(\|B \vee C\|) \cap \|B\| \neq \emptyset$, and (iii) $\text{opt}_{\succeq}(\|A \vee C\|) \cap \|A\| = \emptyset$. From (i), there is some a such that $a \in \text{opt}_{\succeq}(\|A \vee B\|)$ and $a \models A$. From (ii), there is some b such that $b \in \text{opt}_{\succeq}(\|B \vee C\|)$ and $b \models B$. From (iii), $a \notin \text{opt}_{\succeq}(\|A \vee C\|)$. Since $a \models A \vee C$, there is some c such that $c \models A \vee C$ and $a \not\succeq c$. Since $a \in \text{opt}_{\succeq}(\|A \vee B\|)$ and $a \not\succeq c$, $c \not\models A \vee B$, and so $c \not\models A$ and $c \not\models B$. Hence $c \models C$, and so $c \models B \vee C$. Thus, $b \succeq c$. By transitivity of $\succeq$, $a \not\succeq b$. On the other hand, since $a \in \text{opt}_{\succeq}(\|A \vee B\|)$ and $b \models A \vee B$, $a \succeq b$. Contradiction.

For the max rule, it suffices to invoke the above along with Observation 2.6. □

While under the opt rule transitivity is sufficient for ($\gg$-trans), under the max rule it is not:

Observation 2.12. *There is a preference model $M = (W, \succeq, v)$, with $\succeq$ transitive, such that ($\gg$-trans) fails in M under the max rule.*

Proof. Put $M = (W, \succeq, v)$, with $W = \{a, b, c\}$, $\succeq = \{(a, b)\}$ and $v(p) = \{b\}$, $v(q) = \{b, c\}$ and $v(r) = \{a\}$. This is shown in Figure 5. We have $\max_{\succeq}(\|p \vee q\|) = \{b, c\}$, $\max_{\succeq}(\|q \vee r\|) = \{a, c\}$, $\max_{\succeq}(\|p \vee r\|) = \{a\}$, $\|p\| = \{b\}$, $\|q\| = \{b, c\}$ and $\|r\| = \{a\}$. □

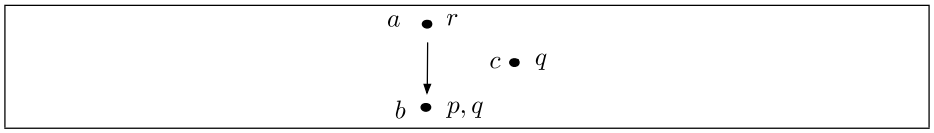


Figure 5: A countermodel to ($\gg$-trans)

(RM) and ($\gg$-trans) are two sample formulas for which the choice between the max rule and the opt rule makes a difference. To get the whole picture (or to get closer to it), we need to extend the scope of our study to examine not individual formulas (chosen randomly) but axiomatic systems. This will be done in Section 4.

2.6 Selection functions

This section provides some background information on so-called selection function semantics. It may seem a distraction from the focus on Hanssonian-type preference semantics. However, this material is needed for subsequent developments, especially in Section 5.1.

Stemming from Stalnaker [1968] and generalized by Chellas [1975], such a semantics was adapted to the present setting by Åqvist [2002].

I call these new structures "selection function models", to distinguish them from those described above. In models of this sort, the betterness relation $\succeq$ is replaced with a so-called selection function $\mathfrak{f}$ from formulas to subsets of W, such that, for all A in $\mathcal{L}$, $\mathfrak{f}(A) \subseteq W$. Intuitively, $\mathfrak{f}(A)$ outputs all the best worlds satisfying A. The evaluation rule for the dyadic obligation operator is phrased thus:

$$M, a \vDash \bigcirc(B/A) \quad \text{iff} \quad \mathfrak{f}(A) \subseteq \|B\|^M$$

From this, on derives the following evaluation rule for permission:

$$M, a \vDash P(B/A) \quad \text{iff} \quad \mathfrak{f}(A) \cap \|B\|^M \neq \emptyset$$

The relevant constraints for $\mathfrak{f}$ are:

(f0) If $\|A\|^M = \|B\|^M$ then $\mathfrak{f}(A) = \mathfrak{f}(B)$ (Syntax-independence)

(f1) $\mathfrak{f}(A) \subseteq \|A\|^M$ (Inclusion)

(f2) $\mathfrak{f}(A) \cap \|B\|^M \subseteq \mathfrak{f}(A \wedge B)$ (Chernoff)

(f3) If $\|A\|^M \neq \emptyset$ then $\mathfrak{f}(A) \neq \emptyset$ (Consistency-preservation)

(f4) If $\mathfrak{f}(A) \subseteq \|B\|^M$ then $\mathfrak{f}(A \wedge B) \subseteq \mathfrak{f}(A)$ (Aizerman)

(f5) If $\mathfrak{f}(A) \cap \|B\|^M \neq \emptyset$ then $\mathfrak{f}(A \wedge B) \subseteq \mathfrak{f}(A) \cap \|B\|^M$ (Arrow)

The reason why these conditions may be regarded as most central will become apparent in Section 3, when moving to the proof theory. Åqvist does not use (f4). It is weaker than (f5) in the following sense.

Fact 2.13. *Given* (f0) *and* (f3)*,* (f5) *implies* (f4)*, but not vice versa (even in the presence of* (f1) *and* (f2)*).*

Proof. Let $\mathfrak{f}(A) \subseteq \|B\|$. Either (i) $\|A\| \neq \emptyset$ or (ii) $\|A\| = \emptyset$. In case (i), $\mathfrak{f}(A) \neq \emptyset$, by (f3). Thus, $\mathfrak{f}(A) \cap \|B\| \neq \emptyset$. (f5) then yields the desired result. In case (ii), $\|A\| = \|A \wedge B\|$. By (f0), $\mathfrak{f}(A) = \mathfrak{f}(A \wedge B)$. So $\mathfrak{f}(A \wedge B) \subseteq \mathfrak{f}(A)$ as required.

To show that the converse implication may fail even in the presence of (f0)-(f3), let $M = (W, \mathfrak{f}, v)$ be such that $W = \{a, b, c\}$, $v(p) = \{a, b\}$ and $v(q) = W$ for all q other than p, and

$$\mathfrak{f}(A) = \begin{cases} \{a, c\} & \text{if } \|A\| = W \\ \|A\| & \text{otherwise} \end{cases}$$

(f0), (f1), (f2) and (f3) hold, and so does (f4). But (f5) fails:

$$\mathfrak{f}(q \wedge p) = \{a, b\} \not\subseteq \mathfrak{f}(q) \cap \|p\| = \{a\} \neq \emptyset$$

This concludes the proof. □

The names used for the first four constraints are from Parent [2015]. All these constraints have known counterparts within the framework of rational choice theory (for an overview, see Moulin [1985]). (f2) is identical to so-called Chernoff's [1954] condition also known as Sen's condition α. (f4) may be regarded as a reformulation of the condition called "Aizerman" in Moulin [1985] and in Lindström [1991]. Therefore it will henceforth be referred to as the Aizerman condition. Strictly speaking, this one is:

$$(\mathfrak{f}4^\star) \qquad \text{If } \mathfrak{f}(A) \subseteq \|B\| \subseteq \|A\| \text{ then } \mathfrak{f}(B) \subseteq \mathfrak{f}(A)$$

It is not difficult to see that, given (f0) and (f1), (f4*) and (f4) are equivalent.

Fact 2.14. *Given (f0) and (f1), (f4*) and (f4) are equivalent.*

Proof. I first verify that, given (f1), (f4*) implies (f4). Assume $\mathfrak{f}(A) \subseteq \|B\|$. By (f1), $\mathfrak{f}(A) \subseteq \|A\| \cap \|B\| = \|A \wedge B\| \subseteq \|A\|$. By (f4*), $\mathfrak{f}(A \wedge B) \subseteq \mathfrak{f}(A)$, as required. For the converse implication, let $\mathfrak{f}(A) \subseteq \|B\| \subseteq \|A\|$. On the one hand, by (f0) $\mathfrak{f}(A \wedge B) = \mathfrak{f}(B)$, since $\|A \wedge B\| = \|B\|$. On the other hand, a direct application of (f4) to $\mathfrak{f}(A) \subseteq \|B\|$ yields $\mathfrak{f}(A \wedge B) \subseteq \mathfrak{f}(A)$. Putting the two together, one gets $\mathfrak{f}(B) \subseteq \mathfrak{f}(A)$ as required. □

(f5) may similarly be regarded as a reformulation of the condition which Hansson [1968] calls Arrow, and so I will henceforth refer to it as the Arrow condition. Strictly speaking, this one is:

$$(\mathfrak{f}5^\star) \qquad \text{If } \|A\| \subseteq \|B\| \text{ and } \mathfrak{f}(B) \cap \|A\| \neq \emptyset \text{ then } \mathfrak{f}(A) = \mathfrak{f}(B) \cap \|A\|$$

Fact 2.15. *Given (f0)-(f3), (f5*) and (f5) are equivalent.*

Proof. See Hansen [1998]. □

Not much more will be needed later about selection functions.

3 Proof systems

This section presents the proof systems to be studied in this chapter.

3.1 Mixed alethic-deontic logics

I will primarily be concerned with four mixed alethic-deontic logics of increasing strength: **E**, **F**, **F**+(CM) and **G**. Systems **E**, **F** and **G** are from Åqvist [1987; 2002]. They correspond to his reconstruction of Hansson [1969]'s system DSDL1, DSDL2 and DSDL3, respectively. **F**+(CM) is from Parent [2014]. The list of all the relevant axioms is given below. For some of the axioms, I introduce special labels in order to facilitate reference to them later on.

The notions of theoremhood, deducibility and consistency (with respect to a given system) are defined as usual. I write $\vdash A$ if A is provable, and $\Gamma \vdash A$ if A is derivable from Γ, where Γ is a set of wffs.

System **E** is defined by the following axioms and rules:

Any axiomatization of classical propositional logic	(PL)
S5-schemata for $\Box$	(S5)
$\bigcirc (B \rightarrow C/A) \rightarrow (\bigcirc(B/A) \rightarrow \bigcirc(C/A))$	(COK)
$\bigcirc (B/A) \rightarrow \Box \bigcirc (B/A)$	(Abs)
$\Box A \rightarrow \bigcirc(A/B)$	(Nec)
$\Box(A \leftrightarrow B) \rightarrow (\bigcirc(C/A) \leftrightarrow \bigcirc(C/B))$	(Ext)
$\bigcirc (A/A)$	(Id)
$\bigcirc (C/A \wedge B) \rightarrow \bigcirc(B \rightarrow C/A)$	(Sh)
If $\vdash A$ and $\vdash A \rightarrow B$ then $\vdash B$	(MP)
If $\vdash A$ then $\vdash \Box A$	(N)

The abbreviations (PL), (S5), (MP) and (N) are self-explanatory. (COK) is the conditional analogue of the familiar distribution axiom K. (Abs) is the absoluteness axiom of [Lewis, 1973], and reflects the fact that the ranking is not world-relative. (Nec) is the deontic counterpart of the familiar necessitation rule. (Ext) permits the replacement of necessarily equivalent sentences in the antecedent of deontic conditionals. (Id) is the deontic analogue of the identity principle. (Sh) is named after Shoham [1988, p. 77], who seems to have been the first to discuss it. The question of whether (Id) is a reasonable law for deontic conditionals has been much debated. A defence of (Id) can be found in

Hansson [1969] and Prakken and Sergot [1997]–this line of defence is discussed in Parent [2012]. (For a different diagnosis, see also Spohn [1975], Makinson [1993], Alchourrón [1993] and Parent [2001].)

For future reference I introduce the following derived principles:

$$\text{If } \vdash A \leftrightarrow B \text{ then } \vdash \bigcirc(C/A) \leftrightarrow \bigcirc(C/B) \qquad \text{(LLE)}$$
$$\text{If } \vdash B \rightarrow C \text{ then } \vdash \bigcirc(B/A) \rightarrow \bigcirc(C/A) \qquad \text{(RW)}$$
$$\bigcirc(B/A) \wedge \bigcirc(C/A) \rightarrow (\bigcirc(B \wedge C/A) \qquad \text{(AND)}$$
$$\bigcirc(C/A) \wedge \bigcirc(C/B) \rightarrow (\bigcirc(C/A \vee B) \qquad \text{(OR)}$$
$$\bigcirc(C/A) \wedge \bigcirc(D/B) \rightarrow \bigcirc(C \vee D/A \vee B) \qquad \text{(OR}'\text{)}$$

The labels (LLE) and (RW) are borrowed from the non-monotonic logic literature. (LLE) and (RW) are mnemonic for "Left Logical Equivalence" and "Right Weakening", respectively.

Theorem 3.1. *(LLE), (RW), (AND), (OR) and (OR′) are derivable in system* **E***.*

Proof. The proofs of (LLE) and (RW) are straightforward, and left to the reader.

For (AND), assume $\bigcirc(B/A)$ and $\bigcirc(C/A)$. From the first, one gets $\bigcirc(C \rightarrow (B \wedge C)/A)$ by (RW). (COK) gives $\bigcirc(C/A) \rightarrow \bigcirc(B \wedge C/A)$. From this and the second hypothesis, one gets $\bigcirc(B \wedge C/A)$.

For (OR), assume $\bigcirc(C/A)$ and $\bigcirc(C/B)$. Using (Ext), one gets $\bigcirc(C/(A \vee B) \wedge A)$ and $\bigcirc(C/(A \vee B) \wedge B)$. By (Sh), $\bigcirc(A \rightarrow C/A \vee B)$ and $\bigcirc(B \rightarrow C/A \vee B)$. By (AND), $\bigcirc((A \rightarrow C) \wedge (B \rightarrow C)/A \vee B)$. By (RW), $\bigcirc((A \vee B) \rightarrow C/A \vee B)$. By (Id), $\bigcirc(A \vee B/A \vee B)$. By (COK), one then gets $\bigcirc(C/A \vee B)$.

(OR′) is easily derived using (OR) and (RW). □

Theorems 3.1 and 3.2 tell us that **E** is equivalently axiomatized by replacing, in **E**, (COK) and (Sh) with (RW), (AND) and (OR).

Theorem 3.2. *(COK) is derivable from (RW) and (AND). (Sh) is derivable from (RW), (Id), (OR) and (LLE).*

Proof. For (COK), assume $\bigcirc(B \rightarrow C/A)$ and $\bigcirc(B/A)$. By (AND), $\bigcirc((B \rightarrow C) \wedge B/A)$. By (RW), $\bigcirc(C/A)$.

For (Sh), suppose $\bigcirc(C/A \wedge B)$. By (RW), $\bigcirc(B \rightarrow C/A \wedge B)$. By (Id) and (RW), $\bigcirc(B \rightarrow C/A \wedge \neg B)$. By (OR) and (LLE), $\bigcirc(B \rightarrow C/A)$. □

The basis of **F** is that of **E** with the single extra axiom:

$$\Diamond A \rightarrow (\bigcirc(B/A) \rightarrow P(B/A)) \qquad \text{(D}^\star\text{)}$$

(D*) is the conditional analogue of the familiar axiom D. Its import is simply that conflicts of obligations are ruled out, for possible antecedents.

F+(CM) and **G** are obtained by supplementing **F** with (CM) and (Sp), respectively:

$$(\bigcirc(B/A) \wedge \bigcirc(C/A)) \rightarrow \bigcirc(C/A \wedge B) \qquad \text{(CM)}$$
$$(P(B/A) \wedge \bigcirc(B \rightarrow C/A)) \rightarrow \bigcirc(C/A \wedge B) \qquad \text{(Sp)}$$

(CM) is the principle of cautious monotony from the non-monotonic logic literature.[12] It can be shown that (CM) and (D*) are independent of each other, given the other axioms of **F**. This is why their addition is considered separately of one another. In the presence of (CM), the following two principles are derivable:

$$\bigcirc(B/A) \wedge \bigcirc(A/B) \rightarrow (\bigcirc(C/A) \leftrightarrow \bigcirc(C/B)) \qquad \text{(CSO)}$$
$$\bigcirc(A/A \vee B) \wedge \bigcirc(B/B \vee C) \rightarrow \bigcirc(A/A \vee C) \qquad (\geq\text{-trans})$$

(CSO) is familiar from the literature on conditional logic. It says that two "deontically" equivalent states of affairs trigger the same obligations. And ($\geq$-trans) expresses a principle of transitivity for a weak notion of preference defined by $A \geq B$ iff $\bigcirc(A/A \vee B)$.[13]

As mentioned, (Sp)−the distinctive axiom of **G**−is equivalent to the principle of rational monotony (RM):[14]

$$(P(B/A) \wedge \bigcirc(C/A)) \rightarrow \bigcirc(C/A \wedge B) \qquad \text{(RM)}$$

F+(CM) is strictly included in **G**, because (CM) is derivable in **G**, but (Sp) is not derivable in **F**+(CM).

Theorem 3.3.

(i) (CM) and (D*) are independent, given the other axioms of **F***;*
(ii) (CSO) is a theorem of **F***+(CM);*
(iii) ($\geq$-trans) is a theorem of **F***+(CM);*
(iv) (Sp) and (RM) are inter-derivable in **E***;*
(v) (CM) is a theorem of **G***;*
(vi) (Sp) is not a theorem of **F***+(CM).*

Proof. The proof of (i) may be found in [Parent, 2014, Section 2.5].

[12] Cf. [Kraus *et al.*, 1990].
[13] Cf. [Kraus *et al.*, 1990, p. 194].
[14] Cf. Section 2.5.

For (ii), assume $\bigcirc(B/A)$, $\bigcirc(A/B)$ and $\bigcirc(C/A)$. From the first and third assumptions, $\bigcirc(C/A \wedge B)$, by (CM). This is equivalent to $\bigcirc(C/B \wedge A)$ by (Ext). Using (Sh), $\bigcirc(A \to C/B)$. From this together with the second assumption, one then gets $\bigcirc(C/B)$, by (RW). For the derivation of $\bigcirc(C/A)$ from $\bigcirc(C/B)$, the argument is similar. This establishes (CSO).

For (iii), assume $\bigcirc(A/A \vee B)$ and $\bigcirc(B/B \vee C)$. Using (OR$'$) and (Ext), $\bigcirc(A \vee B/A \vee B \vee C)$. By (Id) and (RW), $\bigcirc(A \vee B \vee C/A \vee B)$ is a theorem. Using (CSO), one immediately gets $\bigcirc(A/A \vee B \vee C)$. By (Id), $\bigcirc(C/C)$. By (OR$'$) and (Ext), one gets $\bigcirc(A \vee C/A \vee B \vee C)$. By (CM), $\bigcirc(A/(A \vee B \vee C) \wedge (A \vee C))$. By (Ext), $\bigcirc(A/A \vee C)$.

For (iv), suppose $P(B/A)$ and $\bigcirc(C/A)$. By (RW), $\bigcirc(B \to C/A)$, and so $\bigcirc(C/A \wedge B)$ by (Sp). Conversely, suppose $P(B/A)$ and $\bigcirc(B \to C/A)$. By (RM), $\bigcirc(B \to C/A \wedge B)$. Hence $\bigcirc(B \wedge (B \to C)/A)$ by (Sh). One then gets $\bigcirc(C/A)$ by (RW).

For (v). Suppose $\bigcirc(B/A)$ and $\bigcirc(C/A)$. Either $\Diamond A$ or $\neg\Diamond A$. In the first case, $P(B/A)$ by (D*), and so $\bigcirc(C/A \wedge B)$ by (RM). In the second case, $\Box(A \leftrightarrow (A \wedge B))$, and thus $\bigcirc(C/A \wedge B)$ by (Ext). Either way, $\bigcirc(C/A \wedge B)$.

The proof of (vi) is given in Appendix A, where I make use of an observation which will be available only later. □

3.2 Pure deontic conditional logics

The above systems are mixed alethic-deontic logics. Goble [2015, p. 94] shows that each of **F**, **F**+(CM) and **G** has a "pure deontic conditional" counterpart. I borrow this term from Alchourrón [1995, p. 87], who uses the term "pure conditional axiomatisation" to refer to an axiomatisation in a language in which we only have the conditional (obligation) operator as a primitive connective added to those of classical propositional logic. This language still allows iterated modalities and mixed formulas, and thus is still distinct from the language of Hansson's systems.

The key point is that in systems **F**, **F**+(CM) and **G**, the alethic operators $\Box$ and $\Diamond$ become superfluous, because $\Box A$ and $\Diamond A$ turn out to be equivalent with $\bigcirc(\bot/\neg A)$ and $P(\top/A)$, respectively. (This is not the case in **E**, and this is why it is left out of the picture.) Thus, in the description of the three systems, one might eliminate all occurrences of $\Box$ and $\Diamond$ using these definitions, so that everything is written using the deontic modalities only. Drawing on this idea Goble defines three systems, called DDL-D-3, DDL-D-4 and DDL-D-5, using a language with no other primitive modality than $\bigcirc(-/-)$. (Nevertheless, to avoid

cumbersome notation $\Box$ and $\Diamond$ are kept in the language as derived connectives.)[15] The distinctive axiom of DDL-D-4 is (CM), while that of DDL-D-5 is (RM). Roughly speaking, DDL-D-3 may be described as the system that results from **F** by leaving out (D*) (its pure conditional counterpart follows from the other axioms), and by replacing all occurrences of $\Box$ and $\Diamond$ by their definition throughout in (S5), (Abs), (Nec), (Ext) and (N). Goble's own axiomatic characterisation of DDL-D-3 is as follows:

Any axiomatization of classical propositional logic	(PL)
$\Box A \rightarrow A$ (*aka* $\bigcirc(\bot/\neg A) \rightarrow A$)	(T)
If $\vdash A \leftrightarrow B$ then $\vdash \bigcirc(C/A) \leftrightarrow \bigcirc(C/B)$	(LLE)
If $\vdash B \rightarrow C$ then $\vdash \bigcirc(B/A) \rightarrow \bigcirc(C/A)$	(RW)
$\bigcirc(B/A) \wedge \bigcirc(C/A) \rightarrow (\bigcirc(B \wedge C/A)$	(AND)
$\bigcirc(B/A) \wedge \bigcirc(B/C) \rightarrow (\bigcirc(B/A \vee C)$	(OR)
$\bigcirc(A/A)$	(Id)
$\bigcirc(B/A) \rightarrow \bigcirc(\bigcirc(B/A)/C)$	(D$\bigcirc$4)
$P(B/A) \rightarrow \bigcirc(P(B/A)/C)$	(D$\bigcirc$5)
If $\vdash A$ and $\vdash A \rightarrow B$ then $\vdash B$	(MP)
If $\vdash A$ then $\vdash \Box A$ (*aka* $\bigcirc(\bot/\neg A)$)	(N$'$)

(D$\bigcirc$4) and (D$\bigcirc$5) are the dyadic generalization of the well-known principles (4) $\bigcirc A \rightarrow \bigcirc\bigcirc A$ and (5) $PA \rightarrow \bigcirc PA$. (T) and (N$'$) are self-explanatory.

Goble writes that "DDL-D-3 is equivalent to **F**, DDL-D-4 to **F**+(CM) and DDL-D-5 to **G**" [Goble, 2015, p. 102]. All the axioms and rules of each member of the DDL-D family are derivable in the corresponding mixed alethic-deontic logic. Hence the inclusions:

$$\text{DDL-D-3} \subseteq \mathbf{F} \qquad \text{DDL-D-4} \subseteq \mathbf{F}{+}(\text{CM}) \qquad \text{DDL-D-5} \subseteq \mathbf{G}$$

The converse inclusions also hold insofar as $\Box$ and $\Diamond$ are kept as derived connectives in the language of the pure deontic logics, and identified with those appearing in the language of the corresponding mixed alethic-deontic logics. The initial goal was to identify the pure conditional counterparts of Åqvist's systems. For the sake of consistency, one

[15] If in Goble's manuscript we look more closely at the two pairs of operators, we see a subtle difference in notation between them—Åqvist's operators are written as "$\Box$" and "$\Diamond$", while Goble's operators are written as "□" and "◇". For simplicity's sake I will use the same notation for both pairs.

may prefer not to have $\Box$ and $\Diamond$ in the language of the pure deontic logics as a derived connective. In that case, the relationship between the two families of systems should be described differently. One suggestion is to say that each of **F**, **F**+(CM) and **G** can faithfully be embedded into their counterpart in the DDL-D family. That is: there is a translation $\star$ from the language of **F**, **F**+(CM) and **G** into the language of DDL-D-3, DDL-D-4 and DDL-D-5, such that $\star$ preserves both theoremhood and unprovability.

Figure 6 provides a map of the systems I have discussed. An arrow indicates (proper) containment in the sense that the system from which the arrow starts contains all the theorems of the system at which the arrow points, but not vice versa. The systems to the left of the dashed line are mixed alethic-deontic logics, while those to its right are pure deontic logics.

One can find more systems in the literature. In particular, there are also Hansson's DSDL1-3 as axiomatized by Goble [2019], or Lewis's system VN of [1973], which turns out to be equivalent with van Fraassen's system CD of [1972] and Goble's system SDDL of [2003]. However, none will be a part of the discussion. I mentioned that **F** and **G** were meant to be a reconstruction of Hansson's systems DSDL2 and DSDL3. Neither of **F** and **G** contains its DSDL counterpart. Both DSDL2 and DSDL3 have the rule "If $\nvdash \neg A$, then $\vdash P(\top/A)$", while neither of **F** and **G** does.

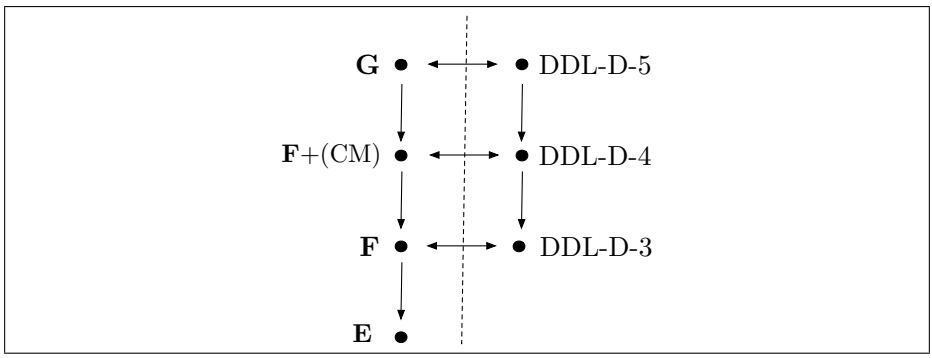


Figure 6: Systems

4 Determination results

This section gives a survey of the determination (*i.e.*, soundness and completeness) results available at the time of writing this chapter. Here I shall be primarily interested in the mixed systems put forth by Åqvist. Two determination results are new. Their proof may be found in the Appendices. To keep this chapter at a reasonable length, the proofs of the other results are omitted. Soundness and completeness are understood in their strong version: they conjointly establish a match between the deductibility and the semantic consequence relations, with no restriction on the cardinality of the premise set Γ. The statement of the theorem is written in the form "$\Gamma \vdash A$ iff $\Gamma \models A$".

4.1 Core results

A synopsis of the core determination results is given in Table 1.

Properties of $\succeq$	max	opt
binary relation	**E**	**E**
limitedness	**F**	**F**
smoothness	**F**+(CM)	**F**+(CM)
smoothness transitivity	**F**+(CM)	**G**

Table 1: Core results

This table must be read as follows. The leftmost column shows the constraints placed on $\succeq$. The top row covers the class of all preference models; one does not require any special properties of $\succeq$ apart from being a relation. The other two columns show the corresponding systems, the middle column for models applying the max rule, and the rightmost one for models applying the opt rule. It is understood that limitedness is defined for max in the max column, and for opt in the opt column.

Below I state formally the results reported in Table 1.

Theorem 4.1.

*(i) Under the opt rule (*resp.*, the max rule),* **E** *is sound and complete with respect to the class of all preference models;*

*(ii) Under the opt rule (*resp.*, the max rule),* **F** *is sound and complete with respect to the class of preference models in which $\succeq$ is opt-limited (resp. max-limited).*

Proof. See [Parent, 2015]. □

Theorem 4.2.

*(i) Under the opt rule (*resp.*, the max rule),* $\mathbf{F}+(CM)$ *is sound and complete with respect to the class of preference models in which* $\succeq$ *is opt-smooth (resp. max-smooth);*

(ii) Under the max rule, $\mathbf{F}+(CM)$ *is sound and complete with respect to the class of preference models in which* $\succeq$ *is max-smooth and transitive.*

Proof. For (i), see [Parent, 2014]. For (ii), see Appendix B. □

Theorem 4.2 (ii) tells us that, under the max rule, and given max-smoothness, the transitivity of $\succeq$ has no import. We will see that this also holds in the absence of max-smoothness. These results are in sharp contrast with those for the opt rule. For instance, in the presence of opt-smoothness, transitivity boosts the logic from **F**+(CM) to **G**.

Theorem 4.3. *Under the opt rule,* **G** *is sound and complete with respect to the class of preference models in which* $\succeq$ *is opt-smooth and transitive.*

Proof. See [Parent, 2014; Parent, 2008]. □

4.2 Adding reflexivity and totalness

Table 2 shows what happens when the constraints of reflexivity and of totalness are added. Reflexivity has no import. Totalness makes a difference only under the max rule in one case, when it is combined with transitivity and smoothness. Below I state formally the results shown in the table.

Theorem 4.4.

*(i) Under the opt rule (*resp.*, the max rule),* **E** *is sound and complete with respect to:*

(a) the class of preference models in which $\succeq$ *is reflexive;*

(b) the class of preference models in which $\succeq$ *is total.*

*(ii) Under the opt rule (*resp.*, the max rule),* **F** *is sound and complete with respect to:*

(a) the class of preference models in which $\succeq$ *is opt-limited (*resp.*, max-limited) and reflexive;*

(b) the class of preference models in which $\succeq$ *is opt-limited (*resp.*, max-limited) and total.*

Proof. See [Parent, 2015]. □

Theorem 4.5.

*(i) Under the opt rule (*resp.*, the max rule),* $\mathbf{F} + (CM)$ *is sound and complete with respect to:*

(a) the class of preference models in which $\succeq$ *is opt-smooth (*resp.*, max-smooth) and reflexive;*

(b) the class of preference models in which $\succeq$ *is opt-smooth (*resp.*, max-smooth) and total.*

(ii) Under the max rule, $\mathbf{F} + (CM)$ *is sound and complete with respect to the class of preference models in which* $\succeq$ *is max-smooth, transitive and reflexive.*

Proof. For (i), see [Parent, 2014]. For (ii), see Appendix B. □

Properties of $\succeq$	max	opt
reflexivity	**E**	**E**
totalness	**E**	**E**
limitedness reflexivity	**F**	**F**
limitedness totalness	**F**	**F**
smoothness reflexivity	**F**+(CM)	**F**+(CM)
smoothness totalness	**F**+(CM)	**F**+(CM)
smoothness transitivity reflexivity	**F**+(CM)	**G**
smoothness transitivity totalness	**G**	**G**

Table 2: Adding reflexivity and totalness

Theorem 4.6.

(i) Under the opt rule, **G** *is sound and complete with respect to:*

(a) the class of preference models in which $\succeq$ *is opt-smooth, transitive and reflexive;*

(b) the class of preference models in which $\succeq$ is opt-smooth, transitive and total.

(ii) Under the max rule, **G** *is sound and complete with respect to the class of preference models in which $\succeq$ is max-smooth, transitive and total.*

Proof. See [Parent, 2014]. □

4.3 Transitivity without smoothness (max rule)

This section reports two determination results for the transitive (and not necessarily smooth) case.

Theorem 4.7. *Under the max rule,* **E** *is sound and complete with respect to:*

(i) the class of preference models in which $\succeq$ is transitive;

(ii) the class of preference models in which $\succeq$ is transitive and reflexive.

Proof. See Appendix C. □

Theorem 4.8. *Under the max rule,* **F** *is sound and complete with respect to:*

(i) the class of preference models in which $\succeq$ is max-limited and transitive;

(ii) the class of preference models in which $\succeq$ is max-limited, transitive and reflexive.

Proof. See Appendix C. □

I summarize these results in Table 3.

Properties of $\succeq$	max	opt
transitivity	**E**	?
transitivity reflexivity	**E**	?
transitivity limitedness	**F**	**G**
transitivity limitedness reflexivity	**F**	**G**

Table 3: Non-smooth transitive betterness under the max rule

The middle column tells us that, under the max rule, transitivity alone has no import, be it combined or not with reflexivity, and be it combined or not with limitedness. This observation does not carry over to the opt rule. Transitivity combined with opt-limitedness boosts the logic from **F**+(CM) to **G**. (Given transitivity, opt-limitedness and opt-smoothness are equivalent.) On the other hand, consider (≫-trans):

$$P(A/A \vee B) \wedge P(B/B \vee C) \rightarrow P(A/A \vee C) \qquad (\gg\text{-trans})$$

We know that under the opt-rule (≫-trans) is valid if $\succeq$ is required to be transitive (cf. Observation 2.11). Thus, under the opt rule, the system obtained by supplementing **E** with (Sp) and (≫-trans) is sound with respect to the class of preference models in which $\succeq$ is transitive and with respect to the class of those in which it is also reflexive. It is not known whether it is also complete with respect to these two classes of models.[16] This is indicated by a question mark in Table 3.

In Section 2.3, I mentioned the possibility of defining "best" in terms of maximization under the transitive closure $\succeq^{\star}$ of $\succeq$. I called this rule of interpretation the quasi-max rule. One has:

Theorem 4.9. *Under the quasi-max rule,* **E** *is sound and complete with respect to:*

(i) the class of all preference models;

(ii) the class of preference models in which $\succeq$ is reflexive.

Proof. This follows from Theorem 4.7 and Observation 2.6. □

4.4 Pure deontic conditional counterparts

Analogous results have been obtained by Goble for his pure deontic systems DDL-D-3, DDL-D-4 and DDL-D-5. Table 4 summarizes these results. As far as the contrast between maximality and optimality is concerned, the story seems to remain the same. I shall make two comments.

- First, there is no known determination result for (i) the class of all preference models (ii) the class of those in which $\succeq$ is required to be reflexive, and (iii) the class of those in which $\succeq$ is required to be total. Hence the presence of a question mark in the relevant cells.

[16] This is also pointed out by [Goble, 2015].

- Second, the axiomatic counterpart of the limitedness assumption changes. In Åqvist's systems, limitedness corresponds to (D*), whose pure deontic conditional counterpart is a theorem of DDL-D-2. The limitedness assumption validates the (T) axiom; this one takes over the role of (D*):

$$\Box A \to A \ (aka \bigcirc (\bot/\neg A) \to A) \qquad \text{(T)}$$

Properties of $\succeq$	max	opt
Binary relation	?	?
reflexivity	?	?
totalness	?	?
limitedness	DDL-D-3	DDL-D-3
limitedness reflexivity	DDL-D-3	DDL-D-3
limitedness totalness	DDL-D-3	DDL-D-3
smoothness	DDL-D-4	DDL-D-4
smoothness reflexivity	DDL-D-4	DDL-D-4
smoothness totalness	DDL-D-4	DDL-D-4
smoothness transitivity	DDL-D-4	DDL-D-5
smoothness transitivity reflexivity	DDL-D-4	DDL-D-5
smoothness transitivity totalness	DDL-D-5	DDL-D-5

Table 4: Pure deontic conditional counterparts

4.5 Methods for proving completeness

Some remarks on the methods for proving the completeness part of the above determination results are in order. They will help the reader to get a feeling of what is involved.

4.5.1 Direct canonical model construction

All the proofs of completeness mentioned above are based on canonical models (see, for instance, [Chellas, 1980]). The proofs of completeness of **F**+(CM) and **G** in [Parent, 2008; Parent, 2014] use a direct canonical model construction. Adapting the canonical model technique to a preference-based setting is not as straightforward as might seem at first sight. Roughly speaking, the worlds in a canonical model are maximal consistent sets (MCSs) of sentences. The main difficulty is to define the comparative goodness relation in such a way that the semantic truth conditions for formulas starting with a deontic operator coincide with the set-membership relation between formulas and maximal consistent sets. In [Åqvist, 1987; Åqvist, 2002], the technique of so-called systematic frame constants is used to define the betterness relation part of the canonical model of **G**. Hansen [1999, p. 130] has shown that the method fails with respect to strong completeness.

For **F**+(CM) and **G**, one can think of suitable constructions. I start with **G**. The basic idea is to work with a point-generated canonical model. The set of all the MCSs is denoted by Ω. Where a is a MCS, a^A denotes $\{B : \bigcirc(B/A) \in a\}$.

Definition 4.10 (Canonical model, **G**)**.** *Let w be a fixed element of Ω. The canonical model generated by w is the structure $M^w = (W, \succeq, V)$ defined by*

(i) $W = \{a \in \Omega : \{A : \Box A \in w\} \subseteq a\}$
(ii) $a \succeq b$ *iff*
  (a) there is no consistent wff A such that $w^A \subseteq b$, or
  (b) there is some $A \in a \cap b$ such that $w^A \subseteq a$
(iii) $v(p) = \{a \in W : p \in a\}$ *for all* $p \in \mathbb{P}$.

Condition (i) says that W is the restriction of Ω to the set of MCSs containing all the wffs A for which $\Box A$ is in the "generating" world w. This is needed to deal with the alethic modalities. The import of condition (ii) is that the best (according to $\succeq$) MCSs among those containing A are precisely those containing all the wffs B for which $\bigcirc(B/A)$ is in the "generating" world w.

The required construction for **F**+(CM) is more complex. The worlds in the universe of the canonical models are not just MCS's, but MCS's labeled with some suitable sentence. This is needed to rank them in terms of goodness. To be more precise, a world becomes a pair whose first element is a MCS a, and whose second element is some formula A

such that $w^A \subseteq a$, where w is the MCS used to generate the canonical model. However, the method also demands that the selected MCS is part of the universe W of the canonical model. Given a MCS w, there may not be any A such that $w^A \subseteq w$. Due to this extra complication, one needs to distinguish between a principal case and a limiting case. I give the full details. For the sake of brevity, $A \geq B$ is used as a shorthand for $\bigcirc(A/A \vee B) \in w$, where w is some MCS.

Definition 4.11 (Canonical model, **F**+(CM), principal case)**.** *Let w be a MCS such that $w^A \subseteq w$ for some A. The canonical model generated by (w, A) is the structure $M^{(w,A)} = (W, \succeq, V)$ defined by*

(i) $W = \{(a, B) : a \in \Omega \ \& \ w^B \subseteq a\}$
(ii) $(a, B) \succeq (b, C)$ *iff: either* $C \not\geq B$ *or* $B \in b$
(iii) $v(p) = \{(a, B) \in W : p \in a\}$ *for all* $p \in \mathbb{P}$.

Definition 4.12 (Canonical model, **F**+(CM), limiting case)**.** *Let w be a MCS such that $w^A \subseteq w$ for no A. Take an arbitrarily chosen wff A. The canonical model generated by (w, A) is the structure $M^{(w,A)} = (W, \succeq, V)$ defined by*

(i) $W = \widetilde{W} \cup \{(w, A)\}$, *where* $\widetilde{W} = \{(a, B) : a \in \Omega \ \& \ w^B \subseteq a\}$
(ii) $\widetilde{\succeq} = \trianglerighteq \cup \{((w, A), (w, A))\} \cup \{(\alpha, (w, A)) : \alpha \in \widetilde{W}\}$ *where* $\trianglerighteq \subseteq \widetilde{W} \times \widetilde{W}$ *is defined as in Definition 4.11, putting* $(a, B) \trianglerighteq (b, C)$ *iff either* $C \not\geq B$ *or* $B \in b$
(iii) $v(p) = \{(a, B) \in W : p \in a\}$ *for all* $p \in \mathbb{P}$.

In [Parent, 2014] these two constructions are used to establish the completeness of **F**+(CM) with respect to the class of models in which $\succeq$ is opt-smooth (*resp.*, max-smooth), Theorem 4.2 (i), and with respect to the class of those in which $\succeq$ is also reflexive or total, Theorem 4.5 (i).

Under the max rule, **F**+(CM) is also sound and complete with respect to the class of models in which $\succeq$ is max-smooth and transitive, and with respect to the class of those in which $\succeq$ is also reflexive. This is Theorem 4.2 (ii) and Theorem 4.5 (ii). These results are new to the aforementioned paper. Their proof is given in Appendix B.

4.5.2 Completeness-via-selection-functions

Contrasting with this direct approach, the method used for **E** and **F** in [Parent, 2015] is indirect, and takes a detour through the alternative modeling in terms of selection functions described in Section 2.6. The proposed approach is related to the two-step methodology used

by [Schlechta, 1997, chap. 2] when discussing representation problems for non-monotonic structures. There are two main steps.

The first step consists in showing soundness and completeness of the systems with respect to appropriate classes of selection function models. I state the needed results in the following theorem, which covers **F**+(CM) and **G** as well.

Theorem 4.13.

(i) **E** *is sound and complete with respect to the class of selection function models* $M = (W, \mathfrak{f}, v)$ *in which* $\mathfrak{f}$ *meets syntax-independence* (ƒ0)*, inclusion* (ƒ1) *and Chernoff* (ƒ2)*;*

(ii) **F** *is sound and complete with respect to the class of selection function models* $M = (W, \mathfrak{f}, v)$ *in which* $\mathfrak{f}$ *meets in addition consistency-preservation* (ƒ3)*;*

(iii) **F** *+(CM) is sound and complete with respect to the class of selection function models* $M = (W, \mathfrak{f}, v)$ *in which* $\mathfrak{f}$ *meets in addition Aizerman* (ƒ4)*;*

(iv) **G** *is sound and complete with respect to the class of selection function models* $M = (W, \mathfrak{f}, v)$ *in which* $\mathfrak{f}$ *meets in addition Arrow* (ƒ5)*.*

Proof. See [Åqvist, 2002, Theorem 77, p. 251]. Let w be a fixed element of Ω. Define the canonical model generated by w as the model $M^w = (W, \mathfrak{f}, v)$ where

- $W = \{a \in \Omega : \{A : \Box A \in w\} \subseteq a\}$
- $\mathfrak{f}(A) = \{a \in \Omega : \{B : \bigcirc(B/A) \in w\} \subseteq a\}$
- $v(p) = \{a \in \Omega : p \in a\}$

Åqvist does not consider (CM). It is a straightforward matter to verify that, in the canonical model for **F**+(CM), $\mathfrak{f}$ meets Aizerman (ƒ4). Details are omitted. □

The second step consists in showing that the selection function semantics is equivalent with the preference-based semantics. One half of the equivalence is relatively easy to establish. This is Theorem 4.14 below. For the reason explained in Section 4, care should be taken with the Arrow condition. Under the max rule it calls for both transitivity and totalness of $\succeq$, while under the opt rule the constraint calls for transitivity only.

Theorem 4.14.

(i) For every preference model $M = (W, \succeq, v)$ *applying the opt rule, there is an equivalent selection function model* $M' = (W, \mathfrak{f}, v)$ *(with*

W and v the same) in which ƒ *meets syntax-independence* (ƒ0)*, inclusion* (ƒ1) *and Chernoff* (ƒ2)*. If* $\succeq$ *is opt-limited, then* ƒ *meets consistency-preservation* (ƒ3)*. If* $\succeq$ *meets opt-smoothness, then* ƒ *meets Aizerman* (ƒ4)*. If* $\succeq$ *is transitive, then* ƒ *meets Arrow* (ƒ5)*.*

(ii) For every preference model $M = (W, \succeq, v)$ *applying the max rule, there is an equivalent selection function model* $M' = (W, \mathfrak{f}, v)$ *(with W and v the same) in which* ƒ *meets syntax-independence* (ƒ0)*, inclusion* (ƒ1) *and Chernoff* (ƒ2)*. If* $\succeq$ *is max-limited, then* ƒ *meets consistency-preservation* (ƒ3)*. If* $\succeq$ *meets max-smoothness, then* ƒ *meets Aizerman* (ƒ4)*. If* $\succeq$ *is transitive and total, then* ƒ *meets Arrow* (ƒ5)*.*

Proof. For (i): starting with $M = (W, \succeq, v)$, define $M' = (W, \mathfrak{f}, v)$ by putting $\mathfrak{f}(A) = \mathrm{opt}_{\succeq}(\|A\|^M)$ for all wff A. For (ii): starting with $M = (W, \succeq, v)$, define $M' = (W, \mathfrak{f}, v)$ by putting $\mathfrak{f}(A) = \max_{\succeq}(\|A\|^M)$ for all wff A. □

The hard part of the proof of equivalence is contained in the following theorem. This one extends a known result from rational choice theory (see, *e.g.*, [Herzberger, 1973]) to the case where the Arrow condition is no longer available.

Theorem 4.15. *For every selection function model* $M = (W, \mathfrak{f}, v)$ *in which* ƒ *meets syntax-independence, inclusion and Chernoff, there is a preference model* $M' = (W', \succeq, v')$ *such that, under the opt rule,* M' *is equivalent to* M*. Furthermore, if* ƒ *meets consistency-preservation, then* $\succeq$ *is opt-limited.*

Proof. See [Parent, 2015, Theorem 3.5]. I recall the proposed construction. Let $M = (W, \mathfrak{f}, v)$. For all $a \in W$, define $\mathcal{Y}_a = \{\|C\|^M \subseteq W \mid a \in \|C\|^M - \mathfrak{f}(C)\}$. Let $\mathcal{Y}_a = \{X_i\}_{i \in I}$. Put $F_a := \prod_{i \in I} X_i$. Intuitively, F_a is the (possibly infinite) cartesian product of all the sets in $\mathcal{Y}_a$. Formally, F_a is the set of all the functions g defined on the index set I such that the value of the function g at a particular index i is an element of X_i:

$$\{g : I \to \bigcup_{i \in I} X_i \mid (\forall i \in I)(g(i) \in X_i)\}$$

The axiom of choice is assumed. Define $M' = (W', \succeq, v')$ as follows:

- $W' = \{\langle a, g\rangle \mid a, b \in W, g \in F_a\}$
- $\langle a, g\rangle \succeq \langle b, g'\rangle$ iff $b \notin \mathrm{Rng}(g))$
- $v'(p) = \{\langle a, g\rangle : a \in v(p)\}$

Rng(g) denotes the range of g, *viz* $\{c \mid \langle i, c\rangle \in g \text{ for some } i \in I\}$. The verification that the construction above actually does the desired job proceeds via a series of lemmas, for which the reader is referred to the above paper. □

Combined with Theorem 4.13 (i) and (ii), Theorem 4.15 yields completeness of **E** with respect to the class of all preference models for the interpretation under the opt rule, and completeness of **F** with respect to the class of those where $\succeq$ is opt-limited under the same interpretation. These two core completeness results carry over to the class of models where $\succeq$ is also total or reflexive, and to the interpretation under the max rule. This is made possible by the following "bridge" result:

Theorem 4.16. *For every preference model $M = (W, \succeq, v)$ in which deontic formulas are interpreted under the opt-rule, there is an equivalent preference model $M = (W', \succeq', v')$ in which $\succeq'$ is total (and hence reflexive), and in which deontic formulas are interpreted under the max-rule (or, equivalently, the opt-rule). Furthermore, if $\succeq$ is opt-limited, then $\succeq'$ is max-limited.*

Proof. See [Parent, 2015, Theorem 3.3]. I recall the proposed construction. Starting with $M = (W, \succeq, v)$, one defines $M' = (W', \succeq', v')$ as follows:

- $W' = \{\langle a, n\rangle \mid a \in W, n \in \mathbb{N}\}$
- $\langle a, n\rangle \succeq' \langle b, m\rangle$ iff $a \succeq b$ or $n \geq m$
- $v'(p) = \{\langle a, n\rangle \mid a \in v(p)\}$

The universe in the output structure is the product set $W \times \mathbb{N}$. Thus, each world a in W has infinitely (albeit countably) many "duplicates" in W'. The order relation on the product set is the lexicographic ordering (or sort of). $\geq$ is total, and so is $\succeq'$. Equivalence between models follows from the fact that the set of optimal elements of $X \subseteq W$ in the input model "matches" the set of maximal elements of $X \times \mathbb{N}$ in the output model, in the sense that:

$$\text{opt}_{\succeq}(X) \times \mathbb{N} = \text{opt}_{\succeq'}(X \times \mathbb{N}) = \max_{\succeq'}(X \times \mathbb{N})$$

This suffices to establish the desired result. □

5 Decidability and automated theorem-proving

5.1 Decidability

The basic result we prove in this section is the decidability of the theoremhood problem "Is A a theorem in such-and-such system?" This will be shown by establishing the so-called finite model property (FMP): any satisfiable formula is satisfiable in a finite model. To simplify matters, this property is shown to hold only with respect to models equipped with a selection function. Decidability of the theoremhood problem in **E**, **F**, **F**+(CM) and **G** follows in the usual way. (See [Chellas, 1980].) The FMP with respect to preference models is put aside.

The FMP with respect to selection function models may be established using the filtration method as adapted by Åqvist [1997; 2000] to a conditional logic setting. I will make a small change to one of his definitions in order to resolve a problem that was pointed out to me by Carmo [2009].

As usual, a model M is said to be finite whenever its universe W is finite. Γ denotes a non-empty and finite set of sentences closed under sub-formulas. § stands for a designated propositional atom in Γ. Put $\top = \S \rightarrow \S$ and $\bot = \neg\top$.

For any selection function model $M = (W, \mathfrak{f}, v)$, the equivalence relation $\sim_\Gamma$ on W is defined by setting

$$a \sim_\Gamma b \text{ iff for every } A \text{ in } \Gamma : a \vDash A \text{ iff } b \vDash A$$

Given $a \in W$, $[a]$ will be the equivalence class of a under $\sim_\Gamma$.

Definition 5.1. *Given some Γ, define the translation function τ, taking every wff into a wff whose propositional atoms are all in Γ, as follows:*

$$\tau(p) = \begin{cases} p & \text{if } p \in \Gamma \\ \S & \text{if } p \notin \Gamma \end{cases}$$

$$\tau(\neg A) = \neg\tau(A) \qquad \tau(A \vee B) = \tau(A) \vee \tau(B)$$
$$\tau(\Box A) = \Box\tau(A) \qquad \tau(\bigcirc(B/A)) = \bigcirc(\tau(B)/\tau(A))$$

Since neither $\top$ nor $\bot$ are primitive symbols, and Γ is non-empty, there is always one such propositional atom § in Γ.

Lemma 5.2. *Let Γ, τ and M be as above. Then, for all wffs A and all $a, b \in W$, if $a \sim_\Gamma b$, then $a \vDash \tau(A)$ iff $b \vDash \tau(A)$.*

Proof. By induction on A. If $A = p$, then either (i) $p \in \Gamma$ or (ii) $p \notin \Gamma$. In case (i), $\tau(p) = p$. In case (ii), $\tau(p) = \S$. In both cases, the claim holds, because $a \sim_\Gamma b$. If $A = B \vee C$ or $A = \neg B$, the result follows directly from the inductive hypothesis. If $A = \Box B$ or $A = \bigcirc(C/B)$, the result follows directly from the evaluation rules for $\Box$ and for $\bigcirc(-/-)$. □

Definition 5.3 (Filtration)**.** *The filtration of* $M = (W, \mathfrak{f}, v)$ *through* Γ *is the model* $M^\star = (W^\star, \mathfrak{f}^\star, v^\star)$ *where:*

(i) $W^\star = \{[a] : a \in W\}$
(ii) $\mathfrak{f}^\star(A) = \{[a] : \exists b \in [a] \;\&\; b \in \mathfrak{f}(\tau(A))\}$
(iii) $v^\star(p) = \{[a] : a \in v(\tau(p))\}$ *for all* $p \in \mathbb{P}$.

Fact 5.4. *(i) If* $a \in W$ *then* $[a] \in W^\star$*; (ii)* $W^\star \neq \emptyset$.

Proof. (i) follows from the reflexivity of $\sim_\Gamma$ and Definition 5.3 (i). (ii) follows from (i) and $W \neq \emptyset$. □

Fact 5.5. $W^\star$ *is finite.*

Proof. The cardinality of $W^\star$ is at most 2^n, where n is the cardinality of Γ. □

A comment on $\mathfrak{f}^\star$ in Definition 5.3 is in order. It is easy to see that $\mathfrak{f}^\star$ is well-defined, in the sense that its definition does not depend upon any particular class representative. That is,

Fact 5.6. *If* $a \sim_\Gamma b$*, then* $[a] \in \mathfrak{f}^\star(A) \leftrightarrow [b] \in \mathfrak{f}^\star(A)$.

Proof. Assume $a \sim_\Gamma b$ and $[a] \in \mathfrak{f}^\star(A)$. It follows that $c \in \mathfrak{f}(\tau(A))$ for some $c \in [a]$. Since $a \sim_\Gamma b$, $c \in [b]$ too, and thus $[b] \in \mathfrak{f}^\star(A)$ as required. For the other direction, the argument is similar. □

Åqvist [1997; 2000] uses the following simpler definition:

$$\mathfrak{f}^\star(A) = \{[a] : a \in \mathfrak{f}(\tau(A))\} \tag{1}$$

Carmo [2009] points out that, if $\mathfrak{f}^\star$ is defined as in (1), then Fact 5.6 may fail as shown in the following example.

Example 5.7. *Put* $M = (W, \mathfrak{f}, v)$ *with* $W = \{a, b\}$*,* $v(p) = W$*, and* $\mathfrak{f}$ *such that*

$$\mathfrak{f}(A) = \begin{cases} \{a\} & \text{if } a \vDash A \\ \|A\|^M & \text{otherwise} \end{cases}$$

$\mathfrak{f}$ *meets syntax-independence, inclusion, Chernoff, consistency-preservation, Aizerman and Arrow. Let* $M^\star = (W^\star, \mathfrak{f}^\star, v^\star)$ *be the filtration of* M *through* $\Gamma = \{p\}$. *We have* $a \sim_{\{p\}} b$. *Assume* $\mathfrak{f}^\star$ *is defined as in (1). We then have* $[a] \in \mathfrak{f}^\star(p)$ *and* $[b] \notin \mathfrak{f}^\star(p)$. *For* $\mathfrak{f}(\tau(p)) = \mathfrak{f}(p) = \{a\}$.

Clause (ii) of Definition 5.3 does not face the above problem. It remains to verify that the entire proof still goes through.

Theorem 5.8 (Filtration Theorem)**.** *Let* Γ, τ, M *and* $M^\star$ *be as above. Then,*

(i) For all wffs A, *if* $A \in \Gamma$, *then* $\tau(A) = A$.
(ii) For all wffs A *and all* $a \in W$:

$$M^\star, [a] \vDash A \text{ iff } M, a \vDash \tau(A).$$

(iii) For all wffs A *in* Γ *and all* $a \in W$:

$$M^\star, [a] \vDash A \text{ iff } M, a \vDash A.$$

Proof. (i) and (ii) are established by induction on A, using the relevant definitions. Clause (iii) is an immediate consequence of (i) and (ii).

For (i), the fact that Γ is closed under subformulas allows us to apply the inductive hypothesis.

I give the full details for (ii) only, focusing on the cases where $A = \Box B$ and $A = \bigcirc(C/B)$, and assuming for induction that the theorem holds for B and C.

- $A = \Box B$. From left-to-right, assume $[a] \vDash \Box B$. By the truth-conditions for $\Box$, we get $[b] \vDash B$ for all $[b]$ in $W^\star$. By the inductive hypothesis, $b \vDash \tau(B)$ for all b in W. Hence $a \vDash \Box\tau(B)$. By definition of τ, $a \vDash \tau(\Box B)$ as required. For the converse direction, argue in reverse.
- $A = \bigcirc(C/B)$. From left-to-right, assume $[a] \vDash \bigcirc(C/B)$, so that $\mathfrak{f}^\star(B) \subseteq \|C\|^{M^\star}$. By definition of τ, to show that $a \vDash \tau(\bigcirc(C/B))$ amounts to showing that $a \vDash \bigcirc(\tau(C)/\tau(B))$. Let $b \in \mathfrak{f}(\tau(B))$. Since $b \in [b]$, $[b] \in \mathfrak{f}^\star(B)$, by Definition 5.3 (ii). Thus, $[b] \vDash C$. By the inductive hypothesis, $b \vDash \tau(C)$, which suffices for $a \vDash \bigcirc(\tau(C)/\tau(B))$. For the other direction, assume $a \vDash \tau(\bigcirc(C/B))$. By definition of τ, $a \vDash \bigcirc(\tau(C)/\tau(B))$. Hence $\mathfrak{f}(\tau(B)) \subseteq \|\tau(C)\|^M$. Let $[b] \in \mathfrak{f}^\star(B)$. By Definition 5.3 (ii), there is some $c \in [b]$ such that $c \in \mathfrak{f}(\tau(B))$. So, $c \vDash \tau(C)$. By Lemma 5.2, $b \vDash \tau(C)$. By the inductive hypothesis, $[b] \vDash C$, which suffices for $[a] \vDash \bigcirc(C/B)$. $\Box$

Theorem 5.9. *Let $M^\star = (W^\star, \mathfrak{f}^\star, v^\star)$ be the filtration of $M = (W, \mathfrak{f}, v)$ through Γ. If $\mathfrak{f}$ meets syntax-independence, inclusion, Chernoff, consistency-preservation, Aizerman or Arrow, then so does $\mathfrak{f}^\star$.*

Proof. This is a matter of running through all the conditions, and showing that they are met.

Syntax-independence. Let $\|A\|^{M^\star} = \|B\|^{M^\star}$. By Theorem 5.8 (ii), $\|\tau(A)\|^M = \|\tau(B)\|^M$. Let $[a] \in \mathfrak{f}^\star(A)$. By Definition 5.3 (ii), $b \in \mathfrak{f}(\tau(A))$ for some $b \in [a]$. Since $\mathfrak{f}$ satisfies syntax-independence, $b \in \mathfrak{f}(\tau(B))$, and hence $[a] \in \mathfrak{f}^\star(B)$. For the other direction, the argument is similar.

Inclusion. Suppose that $[a] \in \mathfrak{f}^\star(A)$. By Definition 5.3 (ii), $b \in \mathfrak{f}(\tau(A))$ for some $b \in [a]$. Since $\mathfrak{f}$ satisfies inclusion, $b \vDash \tau(A)$. By Lemma 5.2, $a \vDash \tau(A)$. By Theorem 5.8 (ii), $[a] \vDash A$.

Chernoff. Suppose that $[a] \in \mathfrak{f}^\star(A) \cap \|B\|^{M^\star}$. By Definition 5.3 (ii), $b \in \mathfrak{f}(\tau(A))$ for some $b \in [a]$. By Theorem 5.8 (ii), $a \vDash \tau(B)$. By Lemma 5.2, $b \vDash \tau(B)$. So, $b \in \mathfrak{f}(\tau(A)) \cap \|\tau(B)\|^M$. Since $\mathfrak{f}$ satisfies Chernoff, $b \in \mathfrak{f}(\tau(A) \wedge \tau(B))$. By definition of τ, $b \in \mathfrak{f}(\tau(A \wedge B))$. By Definition 5.3 (ii), $[a] \in \mathfrak{f}^\star(A \wedge B)$, as required.

Consistency-preservation. Assume $\|A\|^{M^\star} \neq \emptyset$. Hence, there is some $[a] \in W^\star$ such that $[a] \vDash A$. By Theorem 5.8 (ii), $a \vDash \tau(A)$. Since $\mathfrak{f}$ satisfies consistency-preservation, there is $b \in W$ such that $b \in \mathfrak{f}(\tau(A))$. But $b \in [b]$. By Definition 5.3 (ii), $[b] \in \mathfrak{f}^\star(A)$. Hence, $\mathfrak{f}^\star(A) \neq \emptyset$, as required.

Aizerman. Suppose $\mathfrak{f}^\star(A) \subseteq \|B\|^{M^\star}$ and $[a] \in \mathfrak{f}^\star(A \wedge B)$. We need to show that $[a] \in \mathfrak{f}^\star(A)$. By Definition 5.3 (ii) there is some $b \in [a]$ with $b \in \mathfrak{f}(\tau(A \wedge B))$. We show $\mathfrak{f}(\tau(A)) \subseteq \|\tau(B)\|^M$. Let $c \in \mathfrak{f}(\tau(A))$. Since $c \in [c]$, $[c] \in \mathfrak{f}^\star(A)$, Definition 5.3 (ii). By the opening hypothesis, $[c] \vDash B$. By Theorem 5.8 (ii), $c \vDash \tau(B)$, as required. Since $\mathfrak{f}$ satisfies Aizerman, $\mathfrak{f}(\tau(A \wedge B)) \subseteq \mathfrak{f}(\tau(A))$, and thus $b \in \mathfrak{f}(\tau(A))$, which suffices for $[a] \in \mathfrak{f}^\star(A)$, Definition 5.3 (ii).

Arrow. Let $\mathfrak{f}^\star(A) \cap \|B\|^{M^\star} \neq \emptyset$. To show: $\mathfrak{f}^\star(A \wedge B) \subseteq f^\star(A) \cap \|B\|^{M^\star}$. Let $[a] \in \mathfrak{f}^\star(A \wedge B)$. By Definition 5.3 (ii), there is some $b \in [a]$ with $b \in \mathfrak{f}(\tau(A \wedge B))$. By the opening hypothesis, there is some $[c] \in \mathfrak{f}^\star(A)$ with $[c] \vDash B$. By Definition 5.3 (ii), there is some $d \in [c]$ such that $d \in \mathfrak{f}(\tau(A))$. By Theorem 5.8 (ii), $c \vDash \tau(B)$. By Lemma 5.2, $d \vDash \tau(B)$. Hence, $\mathfrak{f}(\tau(A)) \cap \|\tau(B)\|^M \neq \emptyset$. Since $\mathfrak{f}$ meets Arrow, $\mathfrak{f}(\tau(A) \wedge \tau(B)) \subseteq \mathfrak{f}(\tau(A)) \cap \|\tau(B)\|^M$. By definition of τ, $\mathfrak{f}(\tau(A \wedge B)) \subseteq \mathfrak{f}(\tau(A)) \cap \|\tau(B)\|^M$. Hence, $b \in \mathfrak{f}(\tau(A))$ and $b \in \|\tau(B)\|^M$. From the former, $[a] \in \mathfrak{f}^\star(A)$, Definition 5.3 (ii). From the latter, $a \in \|\tau(B)\|^M$, by Lemma 5.2. It follows that

$[a] \vDash B$, Theorem 5.8 (ii). Thus, $\mathfrak{f}^\star(A \wedge B) \subseteq \mathfrak{f}^\star(A) \cap \|B\|^{M^\star}$, as required. □

Theorem 5.10. *The FMP holds with respect to the following classes of selection function models:*

(i) the class of those in which $\mathfrak{f}$ *meets syntax-independence, inclusion and Chernoff;*
(ii) the class of those in which $\mathfrak{f}$ *meets syntax-independence, inclusion, Chernoff, and consistency-preservation;*
(iii) the class of those in which $\mathfrak{f}$ *meets syntax-independence, inclusion Chernoff, consistency-preservation, and Aizerman;*
(iv) the class of those in which $\mathfrak{f}$ *meets syntax-independence, inclusion, Chernoff, consistency-preservation, and Arrow.*

Proof. For (i). Suppose A is satisfiable in some selection function model $M = (W, \mathfrak{f}, v)$ in which $\mathfrak{f}$ meets syntax-independence, inclusion and Chernoff. Thus, there is a world $a \in W$ such that $M, a \vDash A$. Consider the filtration $M^\star = (W^\star, \mathfrak{f}^\star, v^\star)$ of M through the set Γ of all the subformulas of A. By Fact 5.4 (i), $[a] \in W^\star$. By Fact 5.5, $W^\star$ is finite. By Theorem 5.9, $\mathfrak{f}^\star$ meets syntax-independence, inclusion and Chernoff. Furthermore, $A \in \Gamma$. By Theorem 5.8 (iii), $M^\star, [a] \vDash A$. Thus, A is satisfiable in a finite model of the appropriate kind.

For (ii)-(iv), the argument is similar. Details are omitted. □

Since **E**, **F**, **F**+(CM), and **G** are finitely axiomatized, one gets the following spin-off result:

Corollary 5.11. *The theoremhood problem ("Is A a theorem?") in* **E**, **F**, **F***+(CM) and* **G** *is decidable.*

Proof. The argument is standard (see, *e.g.*, [Chellas, 1980]). □

The FMP w.r.t. selection functions is enough to establish the decidability of the theoremhood problem. The question of whether the FMP also holds w.r.t. preference models has an interest in its own right. It is left as a topic for future research.

5.2 Automated theorem proving

This section describes work by [Benzmüller *et al.*, 2019] in automated theorem proving (ATP) for the family of logics discussed in this chapter. Readers who are not interested in automated reasoning can skip this section and go to Section 6.

A specific method called Shallow Semantical Embedding (SSE) is used. The key idea is to use classical higher-order logic (HOL), *i.e.*, Church's type theory [Benzmüller and Andrews, 2019], as a meta-logic in order to represent and model the syntactic and semantic elements of a specific source logic. One can then use off-the-shelf HOL theorem-provers like Isabelle/HOL [Nipkow *et al.*, 2002] or Leo-III [Steen and Benzmüller, 2018; Steen, 2018] for automation. The method was successfully applied to a wide range of non-classical and modal logics—for an overview, see [Benzmüller, 2019] and the references therein. The scope of application of the method has recently been extended to include various prominent deontic logics, including Åqvist's system **E**.[17] The authors focus on the case where deontic formulas are interpreted using the opt rule. It is a straightforward matter to extend the approach to the case where they are interpreted using the max rule, or even an evaluation rule other than one in terms of best, like one of those discussed in Section 6.

In this section I will only briefly describe this work, omitting most of the formal details and proofs, which can be found in the aforementioned paper. The method can be seen as a variant of the so-called standard translation from modal logic to first-order logic [Blackburn *et al.*, 2001]. Possible worlds and propositional letters become individuals and unary predicates, respectively. A distinguished binary predicate symbol r is added to the language of HOL to represent the betterness relation. The modalities are handled by explicit quantification over the set of individuals. One "mimics" the evaluation rules used when evaluating the truth of formulas in a preference model. For example, $\Box A$ translates into:

$$\lambda x.\forall y.Ay$$

And $\bigcirc(B/A)$ translates into:

$$\lambda x\ (\forall y(Ay \wedge (\forall z(Az \rightarrow ryz)) \rightarrow By))$$

This translation holds for the interpretation under the opt rule.

On the HOL side, the following two primitive types are used: i for individuals (or possible worlds); o for the Boolean values. A variant of the standard semantics is used. It is called "generalized" or (after its inventor) "Henkin" semantics. This variant semantics leads to an axiomatizable version of higher-order logic, because the set of functions in a given model need not be complete. (See [Henkin, 1950].)

[17] This is part of a larger project, which aims at mechanizing and automating deontic reasoning. The study [Benzmüller *et al.*, 2020] gives an overview of the project, and documents further the other deontic frameworks covered so far by the SSE method. The Isabelle/HOL theory files are available at `www.logikey.org`.

When working out the formal details, there are three main steps to follow. The first step is to specify an embedding $\lceil\cdot\rceil$, which translates a formula A of **E** into a term $\lceil A\rceil$ of HOL. As mentioned, the clauses of the definition of $\lceil\cdot\rceil$ mirror the evaluation rules used in the semantics of **E**. The second step is to establish that the embedding is sound and complete, that is faithful, in the sense that it preserves both the validity and invalidity of formulas. The establishment of such a result is the main criterion of success. This is Theorem 5.12 below. Intuitively it tells us that under the opt rule a formula A in the language of **E** is valid in the class of all preference models (notation: $\models A$) if and only if the HOL formula $\forall x.\lceil A\rceil x$ is a tautology in every Henkin model (notation: $\models_{\text{HOL}} \forall x.\lceil A\rceil x$).

Theorem 5.12 (Faithfulness of the embedding, [Benzmüller *et al.*, 2019])**.**

$$\models A \textit{ if and only if } \models_{\text{HOL}} \forall x.\lceil A\rceil x$$

Proof. This is [Benzmüller *et al.*, 2019, Theorem 2]. The crux of the argument consists in relating preference models with Henkin models in a truth-preserving way. □

The third and last step consists in encoding the embedding in a concrete theorem-prover like Isabelle/HOL [Nipkow *et al.*, 2002]. Figure 7 displays the encoding obtained for **E**. Some explanations are in order. On line 5, a designated constant "aw" for the actual world is introduced. On lines 28–31, this constant is used to distinguish between global validity (*i.e.*, truth in all worlds) and local validity (*i.e.*, truth at the actual world). On lines 19–26, the dyadic deontic operators are defined by introducing first the notion of optimal A-world.

Here is a (non-exhaustive) list of queries that can be run:

- *Satisfiability:* Is the (finite) set Γ of formulas satisfiable?
- *Validity*: Is formula A valid? Does inference rule R preserve global validity?
- *Entailment:* Does A follow from Γ (with Γ finite)?
- *Correspondance:* Is such-and-such property of the betterness relation sufficient to validate A? Is such-and-such property of the betterness relation necessary to validate A?

When the answer is "no" the model finder Nitpick [Blanchette and Nipkow, 2010] is able to give a counter-example. Similarly, when a formula (or a set of formulas) is satisfiable, Nitpick is able to give a model and a world satisfying the formula (or set of formulas) in question.

```
theory DDLE imports Main
begin
typedecl i (* type for possible worlds *)
type_synonym τ = "(i⇒bool)" (* type for propositions *)
consts aw::i (* actual world *)
consts r :: "i⇒τ"  (infixr "r" 70) (* comparative goodness relation *)
                                   (* Boolean connectives *)
definition ddetop    :: "τ" ("⊤")              where "⊤  ≡ λw. True"
definition ddebot    :: "τ" ("⊥")              where "⊥  ≡ λw. False"
definition ddeneg    :: "τ⇒τ" ("¬_"[52]53)     where "¬A ≡ λw. ¬A(w)"
definition ddeand    :: "τ⇒τ⇒τ" (infixr"∧"51) where "A∧B ≡ λw. A(w)∧B(w)"
definition ddeor     :: "τ⇒τ⇒τ" (infixr"∨"50) where "A∨B ≡ λw. A(w)∨B(w)"
definition ddeimp    :: "τ⇒τ⇒τ" (infixr"→"49) where "A→B ≡ λw. A(w)⟶B(w)"
definition ddeequivt :: "τ⇒τ⇒τ" (infixr"↔"48) where "A↔B ≡ λw. A(w)⟷B(w)"
                                  (* alethic operators *)
definition ddebox  :: "τ⇒τ" ("□") where "□ ≡ λA w.  ∀v. A(v)"
definition ddediomond  :: "τ⇒τ" ("◇") where "◇ ≡ λA w.  ∃v. A(v)"

definition ddeopt :: "τ⇒τ" ("opt<_>") (* deontic operators *)
  where "opt<A> ≡ (λv. ( (A)(v) ∧ (∀x. ((A)(x)  ⟶  v r x)) ) )"
abbreviation(input) msubset :: "τ⇒τ⇒bool" (infix "⊆" 53)
  where "A ⊆ B ≡ ∀x. A x ⟶ B x"
definition ddecond :: "τ⇒τ⇒τ" ("○<_|_>")
  where "○<B|A> ≡ λw. opt<A> ⊆ B"
definition ddeperm :: "τ⇒τ⇒τ" ("P<_|_>")
  where "P<B|A> ≡ ¬○<¬B|A>"

definition ddevalid :: "τ⇒bool" ("⌊_⌋"[8]109) (* global validity *)
  where "⌊p⌋ ≡ ∀w. p w"
definition ddeactual :: "τ⇒bool" ("⌊_⌋ı"[7]105) (* local validity *)
  where "⌊p⌋ı ≡ p(aw)"

end
```

Figure 7: Encoding of system **E** in Isabelle/HOL.

Theorem provers for KLM-style nonmonotonic and conditional logics have been developed, like, e.g., KLMLean 1.0 [Olivetti and Pozzato, 2005], KLM 2.0 [Giordano *et al.*, 2007] and Nescond [Olivetti and Pozzato, 2014]. It would be interesting to compare them with the one described here.

6 Alternative truth-conditions

Despite its length, the chapter does not purport to give an encyclopedic coverage of the field. In this section, I discuss two variant truth-conditions for the conditional obligation operator. As mentioned in the introductory section, more variations are possible. For details, the readers are referred to [Makinson, 1993; Goble, 2015] and references therein.

6.1 The Danielsson-van Fraassen-Lewis truth-conditions

These truth-conditions for deontic sentences are named by Åqvist [1987, p. 199] after their co-inventors: Danielsson [1968], van Fraassen [1972] and Lewis [1973]. One counts $\bigcirc(B/A)$ as true in a world a whenever either there are no A-worlds, or there is some $A \wedge B$-world b such that, as we go up in the ordering, the material implication $A \rightarrow B$ always holds. Hence, all worlds ranked as high as b comply with the obligation in question. This evaluation rule is also used by van Kutschera [1974], Loewer and Belzer [1983] and Goble [2003], among others.

Definition 6.1 ($\exists\forall$ rule)**.** *Given a preference model $M = (W, \succeq, V)$ and a world $a \in W$, we have*

$$M, a \vDash \bigcirc(B/A) \text{ iff } \neg\exists b \ (b \models A) \text{ or } \exists b \ (b \models A \wedge B \ \& \ \forall c \ (c \succeq b \Rightarrow c \models A \rightarrow B)) \quad (\exists\forall)$$

I shall refer to the statement appearing at the right-hand-side of "iff" as the $\exists\forall$ rule. Lewis's preference for the $\exists\forall$ rule is based on his rejection of the limit assumption [1973, p. 98]. The $\exists\forall$ rule handles infinitely ascending chains better than the Hanssonian-type rule in terms of best worlds. Indeed when the A-worlds form an infinitely ascending chain (so that there is no best A-world) under the second rule the formula $\bigcirc(B/A)$ (where B is an arbitrarily chosen formula) becomes (vacuously) true. Thus, when the limit assumption fails, everything is obligatory. With the $\exists\forall$ rule, this is not the case.[18]

Leaving the above issue aside, I now clarify how the $\exists\forall$ rule relates with the opt rule and the max rule.

Theorem 6.2.
(i) The $\exists\forall$ rule implies the opt rule;
(ii) Given totalness of $\succeq$, the $\exists\forall$ rule implies the max rule.

Proof. (ii) follows from (i). To show (i), suppose $\bigcirc(B/A)$ holds at world a in virtue of the $\exists\forall$ rule. This means that either $\neg\exists b \ (b \models A)$ or $\exists b \ (b \models A \wedge B \ \& \ \forall c \ (c \succeq b \Rightarrow c \models A \rightarrow B))$. In the first case, we have $\text{opt}_{\succeq}(\|A\|) = \emptyset$, and so $\text{opt}_{\succeq}(\|A\|) \subseteq \|B\|$. In the second case, consider some $d \in \text{opt}_{\succeq}(\|A\|)$. We have $d \succeq b$ and $d \models A$. So $d \models B$, which suffices for $\text{opt}_{\succeq}(\|A\|) \subseteq \|B\|$ as required. □

[18] Goble's own motivation for using the $\exists\forall$ rule is different. It is not directly related to the limit assumption but to the wish to accommodate conflicts between obligations (see *infra*).

Theorem 6.3.
(i) Given transitivity and opt-limitedness of $\succeq$, the opt rule implies the $\exists\forall$ rule;
(ii) Given transitivity and max-limitedness of $\succeq$, the max rule implies the $\exists\forall$ rule.

Proof. For (i), assume $\text{opt}_{\succeq}(\|A\|) \subseteq \|B\|$. Either (a) $\text{opt}_{\succeq}(\|A\|) = \emptyset$, or (b) $\text{opt}_{\succeq}(\|A\|) \neq \emptyset$. In case (a), by opt-limitedness, $\|A\| = \emptyset$, and so the $\exists\forall$ rule is verified. In case (b), there is some b such that $b \in \text{opt}_{\succeq}(\|A\|)$. We have $b \models B$, by the opening assumption. Let c be such that $c \succeq b$ and $c \models A$. Consider any d such that $d \models A$. We have $b \succeq d$. By transitivity, we then get $c \succeq d$, so that $c \in \text{opt}_{\succeq}(\|A\|)$, and hence $c \models B$, by the opening assumption. Thus, the $\exists\forall$ rule is verified too.
For (ii), the argument is similar. □

The question arises as to how to axiomatize the set of valid formulas for the interpretation under the $\exists\forall$ rule. This question was resolved very early by Lewis and van Fraassen in the case of total orders. Below I recast their result in terms of the systems studied in this chapter. As with Lewis's and van Fraassen's settings, the limit assumption has no impact.

Theorem 6.4. *Under the $\exists\forall$ rule,* **G** *is sound with respect to:*
(i) the class of models in which $\succeq$ is transitive and total (and hence reflexive); and
(ii) the class of models in which $\succeq$ is transitive, total and opt/max-limited (or opt/max-smooth).

Proof. In the presence of transitivity and totalness, opt-limitedness, max-limitedness, opt-smoothness and max-smoothness coincide. All that is required is to show that each axiom of **G** is valid in the class of models in which $\succeq$ is transitive and total, and that the inference rules of **G** preserve validity in this class of models. The argument is routine, and left to the reader. The arguments for (Abs), (Nec), (Ext), (Id) and (Sh) do not call for any of the properties of $\succeq$. (D*) calls for totalness. (Sp) calls for transitivity. (COK) and (CM) call for both totalness and transitivity. For the reader's convenience, I recap these points in the form of a table, Table 5. □

Completeness can be derived from the completeness of **G** under the interpretation applying the opt rule, with respect to the class of models in which $\succeq$ is transitive, total and opt-limited.

Axiom of **G**	Property (or pair of properties) of $\succeq$
(D*)	totalness
(Sp)	transitivity
(COK)	transitivity and totalness
(CM)	transitivity and totalness

Table 5: Axioms and properties under the $\exists\forall$ rule

Theorem 6.5. *Under the $\exists\forall$ rule,* **G** *is complete with respect to:*
(i) the class of models in which $\succeq$ is transitive and total; and
*(ii) the class of models in which $\succeq$ is transitive, total and opt-limited (*resp. *max-limited, opt-smooth and max-smooth).*

Proof. Suppose that $\Gamma \nvdash_{\mathbf{G}} A$. By completeness under the opt rule with respect to the class of models in which $\succeq$ is transitive, total and opt-limited, $\Gamma \not\models A$ over that class of models. By Theorems 6.2 and 6.3, under the $\exists\forall$ rule $\Gamma \not\models A$ over the class of models in which $\succeq$ is transitive, total and opt-limited. Given transitivity and totalness, opt-limitedness, max-limitedness, opt-smoothness and max-smoothness coincide. This establishes (ii). Deleting a constraint on $\succeq$ does not increase the set of semantical consequences. This establishes (i). □

Goble [2003] must be given credit for providing an axiomatization called **DP** in the case of partial orders. In the absence of totalness, (D*), which rules out the possibility of conflicting obligations, goes away. The choice of partial orders may thus be motivated by the need to accommodate conflicts between obligations, these being commonplace.[19] Note that (COK) and (CM) also go away while (Sp) remains. **DP** is a "pure" deontic logic: its language has no other primitive modal operator than $\bigcirc(-/-)$. Furthermore, its semantics uses a betterness relation relativized to worlds, and the truth-conditions make the obligation false when the antecedent is impossible. The proof of completeness for **DP** given by Goble takes a detour through an alternative semantics in terms of multiple preference models. The question as to whether the proof of com-

[19]Here lies Goble's reason for using the $\exists\forall$ rule. With the Hanssonian sort of interpretation, one needs to work with models without the limit assumption ; such models correspond to system **E**. However, **E** contains the following principle of "deontic explosion", $\bigcirc(B/A) \wedge \bigcirc(\neg B/A) \rightarrow \bigcirc(C/A)$, which says that if there is any instance of a deontic dilemma then everything is obligatory. (This is similar to the point made above in relation to the limit assumption, page 49). A survey of the state of the art regarding the treatment of conflicts between obligations may be found in [Goble, 2013].

pleteness for **DP** can be adapted to the present setting is left as a topic for future research. Furthermore, one would like to know what happens within this set-up when transitivity goes away. The question of how to axiomatize the corresponding logic is left as a topic for future research too.

6.2 The Burgess-Boutilier-Lamarre truth-conditions

The evaluation rule used by Burgess [1981], Boutilier [1994] and others has a "$\forall\exists\forall$" structure. This alternative evaluation rule has two technical attractions. First, as noted by Boutilier and independently by Lamarre [1991], it permits the reduction of the dyadic obligation operator to a monadic modal operator. Second, as mentioned by Lewis [1981, p. 230], it enables one to have a fairly strong dyadic deontic logic without committing to either a form of the limit assumption or totalness for $\succeq$. Makinson [1993, p. 346] gives a similar motivation. We see a similar rule in the Kratzer semantics for conditionals (see Kratzer [1991, Definition 13]) and in Veltman [1985]'s logic for conditionals.

Definition 6.6 ($\forall\exists\forall$ rule)**.** *Given a preference model M, and some world a in M, we have*

$$\begin{aligned} a \vDash \bigcirc(B/A) \text{ iff } \forall b \text{ if } b \models A \text{ then} \\ \exists c \text{ s.t. } c \succeq b \;\&\; c \models A \;\&\; \qquad (\forall\exists\forall) \\ \forall d\,(d \succeq c \;\Rightarrow\; d \models A \rightarrow B) \end{aligned}$$

I will refer to the statement at the right-hand side of "iff" as the $\forall\exists\forall$ rule. I just described this rule as a way to avoid commitment to totalness for $\succeq$. This was Lewis's primary motivation. (See also [Kaufmann, 2017, §3].) It is worth mentioning that this benefit comes with a cost: (RM) goes away, while (D*) remains. The argument for (D*) is part of the proof of Theorem 6.10 below. I show the failure of (RM).

Observation 6.7. *There is a preference model $M = (W, \succeq, v)$, with $\succeq$ reflexive and transitive, in which (RM) fails under the $\forall\exists\forall$ rule.*

Proof. Put $M = (W, \succeq, v)$, with $W = \{a, b, c\}$, $\succeq$ the reflexive closure of $\{(b, a), (c, c)\}$ and $v(p) = W$, $v(q) = \{b, c\}$ and $v(r) = \{a, c\}$. This is shown in Figure 8, where reflexivity is left implicit. In this model, $\succeq$ is (vacuously) transitive. We have:

- $a \models \bigcirc(q/p)$
- $a \models \neg \bigcirc (\neg r/p)$ (witness: c)
- $a \not\models \bigcirc(q/p \wedge r)$ (witness: a) □

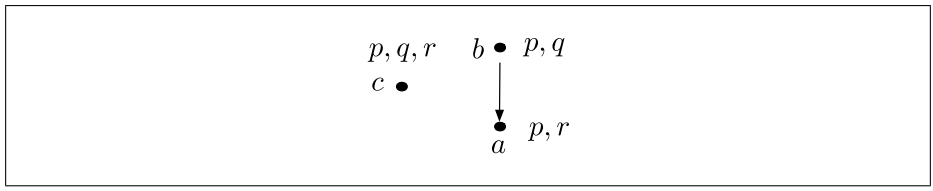


Figure 8: A countermodel to (RM) under the ∀∃∀ rule

Theorem 6.8 clarifies how the ∀∃∀ rule relates with the ∃∀ rule.

Theorem 6.8.

(i) Given reflexivity of $\succeq$*, the* ∀∃∀ *rule implies the* ∃∀ *rule;*

(ii) Given both transitivity and totalness of $\succeq$*, the* ∃∀ *rule implies the* ∀∃∀ *rule.*

Proof. For (i), suppose the ∀∃∀ rule holds, but not the ∃∀ rule. Hence, there is some b_1 such that $b_1 \models A$ and

$$\forall b \ \ (b \models A \wedge B \ \ \Rightarrow \ \ \exists c \, (c \succeq b \ \ \& \ \ c \models A \ \ \& \ \ c \not\models B)) \qquad (\alpha_1)$$

By the ∀∃∀ rule, there is some c_1 such that $c_1 \succeq b_1$, $c_1 \models A$ and

$$\forall d \, (d \succeq c_1 \ \Rightarrow \ d \models A \rightarrow B) \qquad (\alpha_2)$$

By reflexivity, $c_1 \succeq c_1$, and so $c_1 \models B$. By (α_1), there is some d_1 such that $d_1 \succeq c_1$, $d_1 \models A$ and $d_1 \not\models B$. This contradicts (α_2).

For (ii), suppose the ∃∀ rule holds, but not the ∀∃∀ rule. From the latter, there is some b_1 such that $b_1 \models A$ and

$$\forall c \, (c \succeq b_1 \ \& \ c \models A \ \Rightarrow \ \exists d \, (d \succeq c \ \& \ d \models A \ \& \ d \not\models B)) \qquad (\beta_1)$$

For the ∃∀ rule to hold, it must be the case that there is some b_2 such that $b_2 \models A \wedge B$ and

$$\forall c \, (c \succeq b_2 \ \Rightarrow \ c \models A \rightarrow B) \qquad (\beta_2)$$

By totalness, either (a) $b_1 \succeq b_2$ or (b) $b_2 \succeq b_1$. In case (a), (β_2) yields $b_1 \models A \rightarrow B$. By reflexivity of $\succeq$, $b_1 \succeq b_1$. By (β_1), there is some d_1 such that $d_1 \succeq b_1$, $d_1 \models A$ and $d_1 \not\models B$. By transitivity, $d_1 \succeq b_2$, and so by (β_2), $d_1 \models A \rightarrow B$, a contradiction. In case (b), (β_1) yields that there is some d_1 such that $d_1 \succeq b_2$ and $d_1 \models A$ and $d_1 \not\models B$, a result that immediately contradicts (β_2). □

It is noteworthy that, in the presence of the limit assumption, the ∀∃∀ rule coincides with the max rule.

Theorem 6.9.
(i) The $\forall\exists\forall$ rule implies the max rule;
(ii) Given reflexivity, transitivity and max-smoothness of $\succeq$, the max rule implies the $\forall\exists\forall$ rule.

Proof. For (i), suppose the $\forall\exists\forall$ rule holds, and let $b \in \max_{\succeq}(\|A\|)$. Since $b \models A$, there is some c such that $c \succeq b$, $c \models A$ and

$$\forall d\ (d \succeq c\ \Rightarrow\ d \models A \rightarrow B) \qquad (\gamma_1)$$

Since $b \in \max_{\succeq}(\|A\|)$, $b \succeq c$. (γ_1) then yields $b \models B$, which suffices for $\max_{\succeq}(\|A\|) \subseteq \|B\|$.

For (ii), suppose the max rule holds, and let b be such that $b \models A$. By max-smoothness either (a) $b \in \max_{\succeq}(\|A\|)$ or (b) there is c such that $c \succ b$ and $c \in \max_{\succeq}(\|A\|)$. Suppose (a) applies. By reflexivity, $b \succeq b$. Also $b \models A$. Let c be such that $c \succeq b$ and $c \models A$. Let d be such that $d \succeq c$ and $d \models A$. By transitivity of $\succeq$, $d \succeq b$. By maximality of b, $b \succeq d$. By transitivity of $\succeq$ again, $c \succeq d$. Hence, $c \in \max_{\succeq}(\|A\|)$. It then follows that $c \models B$ as required. The argument for (b) is similar, working with c instead of b. □

Theorem 6.10. *Under the $\forall\exists\forall$ rule, $\mathbf{F}$+(CM) is sound with respect to the class of models in which $\succeq$ is reflexive and transitive.*

Proof. This is just a matter of verifying that the axioms of $\mathbf{F}$+(CM) are valid. (Ext) and (Abs) hold independently of the reflexivity and transitivity of $\succeq$. (Nec), (Id) and ($\mathrm{D}^\star$) each call for the reflexivity of $\succeq$. (CM) and (COK) call for transitivity of $\succeq$, while (Sh) calls for both transitivity and reflexivity. For the reader's convenience, I recap these points in the form of a table, Table 6. I give the argument for ($\mathrm{D}^\star$)and (CM) only.

For ($\mathrm{D}^\star$), suppose (i) $a \models \Diamond A$ and (ii) $a \models \bigcirc(B/A)$. To show: $a \models P(B/A)$, *i.e.*, $a \not\models \bigcirc(\neg B/A)$. From (i), there is some b be such that $b \models A$. Let c be such that $c \succeq b$ and $c \models A$. From (ii), there is some $d \succeq c$ such that $d \models A$ and

$$\forall e\ (e \succeq d\ \Rightarrow\ e \models A \rightarrow B) \qquad (\delta_1)$$

By reflexivity, $d \succeq d$, and hence by (δ_1) $d \models B$, *i.e.*, $d \not\models \neg B$. Hence, $a \not\models \bigcirc(\neg B/A)$ as required.

For (CM), suppose (i) $a \models \bigcirc(B/A)$ and (ii) $a \models \bigcirc(C/A)$. Let b_1 be such that $b_1 \models A \wedge B$. By (i), there is some $b_2 \succeq b_1$ such that $b_2 \models A$ and

$$\forall c\ (c \succeq b_2\ \Rightarrow\ c \models A \rightarrow B) \qquad (\delta_2)$$

By (ii), there is some $b_3 \succeq b_2$ such that $b_3 \models A$ and

$$\forall c\,(c \succeq b_3 \;\Rightarrow\; c \models A \to C) \qquad (\delta_3)$$

By (δ_2), $b_3 \models B$ and hence $b_3 \models A \wedge B$. By transitivity of $\succeq$, $b_3 \succeq b_1$. Let d be such that $d \succeq b_3$ and $d \models A \wedge B$. Obviously, $d \models A$. By (δ_3), $d \models C$, which suffices for $a \models \bigcirc(C/A \wedge B)$. □

Axiom of **F**+(CM)	Property (or pair of properties) of $\succeq$
(Nec)	reflexivity
(Id)	reflexivity
(D*)	reflexivity
(CM)	transitivity
(COK)	transitivity
(Sh)	reflexivity and transitivity

Table 6: Axioms and properties under the ∀∃∀ rule

Theorem 6.11. *Under the ∀∃∀ rule,* **F***+(CM) is complete with respect to the class of models in which $\succeq$ is reflexive and transitive.*

Proof. Suppose $\Gamma \not\vdash_{\mathbf{F}+(CM)} A$. By Theorem 4.5 (ii), for the interpretation under the max rule we have that $\Gamma \not\models A$ over the class of models in which $\succeq$ is reflexive, transitive and max-smooth. By Theorem 6.9, the observation that $\Gamma \not\models A$ over the class of models in which $\succeq$ is reflexive, transitive and max-smooth carries over to the interpretation under the ∀∃∀ rule. That $\Gamma \not\models A$ continues to apply, *mutatis mutandis*, with respect to the class of models in which $\succeq$ is only reflexive and transitive. □

As with the ∃∀ rule, the limit assumption has no impact.

Corollary 6.12. *Under the ∀∃∀ rule,* **F***+(CM) is sound and complete with respect to the class of models in which $\succeq$ is reflexive, transitive and max-smooth (*resp. *max-limited).*

Proof. Soundness follows from the fact that no axiom requires max-smoothness or max-limitedness. Completeness with respect to the class of models with max-smoothness has just been established as part of the proof of Theorem 6.11. Completeness with respect to the class of models with max-limitedness follows from this and Observation 2.8 (a) (i). □

It should be pointed out that Theorem 6.11 echoes the axiomatization result obtained by Goble [2014] for the Kratzer conditional.

I end this section by showing that the assumption of totalness boosts the logic from **F**+(CM) to **G**.

Theorem 6.13. *Under the* $\forall\exists\forall$ *rule,* **G** *is sound and complete with respect to:*

(i) the class of models in which $\succeq$ *is transitive and total (and hence reflexive); and*

(ii) the class of models in which $\succeq$ *is transitive, total and max-limited (*resp. *max-smooth, opt-limited and opt-smooth).*

Proof. For soundness, it suffices to verify that (Sp) holds is valid when $\succeq$ is required to be total. Consider a model M and a world a in M such that (i) $a \models P(B/A)$, (ii) $a \models \bigcirc(B \rightarrow C/A)$ and (iii) $a \not\models \bigcirc(C/A \wedge B)$. From (iii), there is some b_1 such that $b_1 \models A \wedge B$ and

$$\forall c\,((c \succeq b_1 \;\&\; c \models A \wedge B) \;\Rightarrow\; \exists d\,(d \succeq c \;\&\; d \models A \wedge B \;\&\; d \not\models C)) \quad (\epsilon_1)$$

From (ii), there is some $b_2 \succeq b_1$ with $b_2 \models A$ and

$$\forall c\,(c \succeq b_2 \;\Rightarrow\; c \models A \rightarrow (B \rightarrow C)) \quad (\epsilon_2)$$

From (i), there is some b_3 such that $b_3 \models A$ and

$$\forall c\,((c \succeq b_3 \;\&\; c \models A) \;\Rightarrow\; \exists d\,(d \succeq c \;\&\; d \models A \wedge B)) \quad (\epsilon_3)$$

By totalness, either (1) $b_2 \succeq b_3$ or (2) $b_3 \succeq b_2$. We argue that, in both cases, there is some b_4 with $b_4 \succeq b_2$ and $b_4 \models A \wedge B$. In case (1), (ϵ_3) immediately yields this result. In case (2), $b_3 \succeq b_3$ by reflexivity, and so (ϵ_3) tells us that there is some b_4 with $b_4 \succeq b_3$ and $b_4 \models A \wedge B$. By transitivity, $b_4 \succeq b_2$. Thus, either way, there is some b_4 with $b_4 \succeq b_2$ and $b_4 \models A \wedge B$. By transitivity, $b_4 \succeq b_1$. (ϵ_1) then yields that there is some b_5 with $b_5 \succeq b_4$ and $b_5 \models A \wedge B \wedge \neg C$. This contradicts (ϵ_2), since $b_5 \succeq b_2$, by transitivity.

Completeness follows at once from Theorems 6.5 and 6.8. □

7 Conclusion

The chapter has provided a survey of results related to the meta-theory of dyadic deontic logics in Hansson's tradition, focusing on axiomatization issues. The goal was to provide a "roadmap" of the different systems that can be obtained, depending on the special properties envisaged for

the betterness relation, and depending on whether "best" means "optimal" or "maximal". Four systems of increasing strength were discussed, and related to (combinations of) properties of the betterness relation. The most remarkable finding in this study is that the contrast between the two notions of "best" is not as significant as one may think,[20] because in an appreciable number of cases the determined logic remains the same no matter which definition is used. Another unexpected outcome is that an apparently strong condition like totalness (and also, sometimes, transitivity) is somewhat idle, because in quite a number of cases its imposition does not affect the logic.

At least two qualifications of these findings are worth noting. First, we have noticed an asymmetry between maximality and optimality in two cases, when transitivity interacts with totalness (and smoothness), and when transitivity is considered alone. The latter case is not fully understood yet because no completeness result for optimality has been reported. Second, the correlations between the properties of the betterness relation and the axioms are not the same when variant truth-conditions for the conditional are used in order to circumvent the limit assumption. Two such variant truth-conditions are the $\exists\forall$ rule and the $\forall\exists\forall$ rule. Under the former a completeness theorem is available for models with a transitive and total relation, and under the latter for models with a reflexive and transitive relation. But we still do not know the full picture. In particular it is not known what happens when transitivity goes away.

For the sake of exhaustiveness, decidability of the theoremhood problem and automated theorem-proving were also discussed. The decidability of the theoremhood problem in the four proof systems studied in this chapter was established, by taking a detour through a modeling in terms of a selection function. Reasoning tasks were automated via a faithful embedding into HOL. These topics have an interest in their own right. However no deeper insight on the above issues was gained. Looking at computational complexity is a natural next step.

References

[Alchourrón, 1993] C. Alchourrón. Philosophical foundations of deontic logic and the logic of defeasible conditionals. In J.-J.Ch. Meyer and R.J. Wieringa,

[20] Bossert and Suzumara (personal communication) reached a similar conclusion within the framework of rational choice theory (cf. [Bossert and Suzumura, 2010, chapter 3].

editors, *Deontic Logic in Computer Science*, pages 43–84. John Wiley & Sons, Inc., New York, 1993.

[Alchourrón, 1995] C. Alchourrón. Defeasible logic: demarcation and affinities. In G. Crocco, L. Fariñas del Cerro, and A. Herzig, editors, *Conditionals: From Philosophy to Computer Science*, pages 67–102. Oxford University Press, Oxford, 1995.

[Åqvist, 1987] L. Åqvist. *An Introduction to Deontic logic and the Theory of Normative Systems*. Bibliopolis, Naples, 1987.

[Åqvist, 1993] L. Åqvist. A completeness theorem in deontic logic with systematic frame constants. *Logique & Analyse*, 36(141-142):177–192, 1993.

[Åqvist, 1997] L. Åqvist. On certain extensions of von Kutschera's preference-based dyadic deontic logic. In W. Lenzen, editor, *Das weite Spektrum der analytischen Philosophie: Festschrift fur Franz von Kutschera*, pages 8–23. Walter de Gruyter, Berlin, New York, 1997.

[Åqvist, 2000] L. Åqvist. Three characterizability problems in deontic logic. In R. Demolombe and R. Hilpinen, editors, *Proceedings of the* 5^{th} *International Workshop on Deontic Logic In Computer Science (*Δ*EON'00)*, pages 16–41. ONERA-DGA, 2000.

[Åqvist, 2002] L. Åqvist. Deontic logic. In D. Gabbay and F. Guenthner, editors, *Handbook of Philosophical Logic*, volume 8, pages 147–264. Kluwer Academic Publishers, Dordrecht, Holland, 2nd edition, 2002. Originally published in [Gabbay and Guenthner, 1984, pp. 605–714].

[Asher and Bonevac, 1997] N. Asher and D. Bonevac. Common sense obligation. In Nute [1997], pages 159–203.

[Benzmüller and Andrews, 2019] C. Benzmüller and P. Andrews. Church's type theory. In E. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, summer 2019 edition, 2019. `https://plato.stanford.edu/entries/type-theory-church/`.

[Benzmüller *et al.*, 2019] C. Benzmüller, A. Farjami, and X. Parent. Åqvist's dyadic deontic logic E in HOL. *Journal of Applied Logics–IfCoLog*, 6:715–732, 2019.

[Benzmüller *et al.*, 2020] C. Benzmüller, X. Parent, and L. van der Torre. Designing normative theories of ethical reasoning: LogiKEy formal framework, methodology, and tool support. *Artificial Intelligence*, 287:103348, 2020.

[Benzmüller, 2019] C. Benzmüller. Universal (meta-)logical reasoning: Recent successes. *Science of Computer Programming*, 172:48–62, 2019.

[Blackburn *et al.*, 2001] P. Blackburn, M. de Rijke, and Y. de Venema. *Modal Logic*. Cambridge Tracts in Theoretical Computer Science. Cambridge University Press, 2001.

[Blanchette and Nipkow, 2010] J. C. Blanchette and T. Nipkow. Nitpick: A counterexample generator for higher-order logic based on a relational model finder. In M. Kaufmann and L. C. Paulson, editors, *Interactive Theorem Proving, First International Conference, ITP 2010, Edinburgh, UK, July 11-14, 2010. Proceedings*, volume 6172 of *Lecture Notes in Computer Science*,

pages 131–146. Springer, 2010.

[Bossert and Suzumura, 2010] W. Bossert and K. Suzumura. *Consistency, Choice, and Rationality*. Harvard University Press, Cambridge, 2010.

[Boutilier, 1994] G. Boutilier. Toward a logic for qualitative decision theory. In J. Doyle, E. Sandewall, and P. Torasso, editors, *Proceedings of the 4th International Conference on Principles of Knowledge Representation and Reasoning (KR-94)*, pages 75–86. Morgan Kaufman, Bonn, 1994.

[Burgess, 1981] J. P. Burgess. Quick completeness proofs for some logics of conditionals. *Notre Dame J. Formal Logic*, 22(1):76–84, 1981.

[Carmo, 2009] J. Carmo. Private communication, 2009.

[Chellas, 1975] B. Chellas. Basic conditional logic. *Journal of Philosophical Logic*, 4(2):133–153, 1975.

[Chellas, 1980] B. Chellas. *Modal Logic*. Cambridge University Press, Cambridge, 1980.

[Chernoff, 1954] H. Chernoff. Rational selection of decision functions. *Econometrica*, 22(4):422–443, 1954.

[Chisholm, 1963] R. Chisholm. Contrary-to-duty imperatives and deontic logic. *Analysis*, 24:33–36, 1963.

[Danielsson, 1968] S. Danielsson. *Preference and Obligation, Studies in the Logic of Ethics*. Filosofiska Färeningen, 1968.

[Fehige, 1994] C. Fehige. The limit assumption in deontic (and prohairetic) logic. In G. Meggle and U. Wessels, editors, *Analyomen 1*, pages 42–56. De Gruyter, Berlin, 1994.

[Gabbay and Guenthner, 1984] D. Gabbay and F. Guenthner, editors. *Handbook of Philosophical Logic*, volume II. Reidel, Dordrecht, Holland, 1st edition, 1984.

[Gabbay *et al.*, 2013] D. Gabbay, J. Horty, X. Parent, R. van der Meyden, and L. van der Torre, editors. *Handbook of Deontic Logic and Normative Systems*, volume 1. College Publications, London, 2013.

[Giordano *et al.*, 2007] L. Giordano, V. Gliozzi, and G. L. Pozzato. KLMLean 2.0: A theorem prover for KLM logics of nonmonotonic reasoning. In N. Olivetti, editor, *Automated Reasoning with Analytic Tableaux and Related Methods*, pages 238–244, Berlin, Heidelberg, 2007. Springer.

[Goble, 2003] L. Goble. Preference semantics for deontic logics. Part I: Simple models. *Logique & Analyse*, 46(183-184):383–418, 2003.

[Goble, 2013] L. Goble. Prima facie norms, normative conflicts and dilemmas. In Gabbay et al. [2013], pages 241–352.

[Goble, 2014] L. Goble. Further notes on Kratzer semantics for modality, with application to dyadic deontic logic, 2014. Unpublished.

[Goble, 2015] L. Goble. Models for dyadic deontic logics, 2015. Unpublished (version dated 8 October 2015).

[Goble, 2019] L. Goble. Axioms for Hansson's dyadic deontic logics. *Filosofiska Notiser*, 6(1):13–61, 2019.

[Goldman, 1977] H. Goldman. David Lewis's semantics for deontic logic. *Mind*, 86(342):242–248, 1977.

[Hansen, 1998] J. Hansen. Notes to my DEON'98 contribution, 1998. Available on-line at the address: `http://www.hh.shuttle.de/win/Joerg.Hansen/Deontic.html` (this document was initially distributed at the DEON conference held in Bologna in 1998, where a first version of [Hansen, 1999] was presented).

[Hansen, 1999] J. Hansen. On relations between Åqvist's deontic system G and van Eck's deontic temporal logic. In P. Mc Namara and H. Prakken, editors, *Norms, Logics and Information Systems*, Frontiers in Artificial Intelligence and Applications, pages 127–144. IOS Press, Amsterdam, 1999.

[Hansen, 2005] J. Hansen. Conflicting imperatives and dyadic deontic logic. *Journal of Applied Logic*, 3(3-4):484–511, 2005.

[Hanson, 1965] W. H. Hanson. Semantics for deontic logic. *Logique & Analyse*, 8:177–190, 1965.

[Hanson, 1969] B. Hanson. An analysis of some deontic logics. *Noûs*, 3(4):373–398, 1969. Reprinted in [Hilpinen, 1971, pp. 121-147].

[Hansson, 1968] B. Hansson. Choice structures and preference relations. *Synthese*, 18(4):443–458, 1968.

[Hansson, 2009] S.O. Hansson. Preference-based choice functions: a generalized approach. *Synthese*, 157, 2009.

[Henkin, 1950] L. Henkin. Completeness in the theory of types. *Journal of Symbolic Logic*, 15(2):81–91, 06 1950.

[Herzberger, 1973] H. Herzberger. Ordinal preference and rational choice. *Econometrica*, 41(2):187–237, 1973.

[Hilpinen and McNamara, 2013] R. Hilpinen and P. McNamara. Deontic logic: a historical survey and introduction. In Gabbay et al. [2013], pages 3–136.

[Hilpinen, 1971] R. Hilpinen, editor. *Deontic Logic: Introductory and Systematic Readings*. Reidel, Dordrecht, 1971.

[Hilpinen, 2001] R. Hilpinen. Deontic logic. In L. Goble, editor, *The Blackwell Guide to Philosophical Logic*, pages 159–182. Blackwell Publishers, Malden, 2001.

[Horty, 2001] J. Horty. *Agency and Deontic Logic*. Oxford University Press, New York, 2001.

[Horty, 2014] J. Horty. Deontic modals: Why abandon the classical semantics? *Pacific Philosophical Quarterly*, 95(4):424–460, 2014.

[Jackson, 1985] F. Jackson. On the semantics and logic of obligation. *Mind*, 94(374):177–195, 1985.

[Kaufmann, 2017] S. Kaufmann. The limit assumption. *Semantics and Pragmatics*, 10:1–29, 2017.

[Kratzer, 1991] A. Kratzer. Modality. In A. von Stechow and D. Wunderlich, editors, *Semantics: An International Handbook of Contemporary Research*, page 639–650. De Gruyter, Berlin, summer 2019 edition, 1991.

[Kratzer, 2012] A. Kratzer. The notional category of modality. In A. Kratzer, editor, *Modals and Conditionals*, pages 27–69. Oxford University Press, 2012.

[Kraus *et al.*, 1990] S. Kraus, D. Lehmann, and M. Magidor. Nonmonotonic reasoning, preferential models and cumulative logics. *Artificial Intelligence*, 44(1-2):167–207, 1990.

[Lamarre, 1991] P. Lamarre. S4 as the conditional logic of nonmonotonicity. In *Proceedings of the Second International Conference on Principles of Knowledge Representation and Reasoning*, pages 357–367. Morgan Kaufmann, 1991.

[Lehmann and Magidor, 1992] D. Lehmann and M. Magidor. What does a conditional knowledge base entail? *Artificial Intelligence*, 55(1):1–60, 1992.

[Lewis, 1973] D. Lewis. *Counterfactuals*. Blackwell, Oxford, 1973.

[Lewis, 1974] D. Lewis. Semantic analyses for dyadic deontic logic. In S. Stenlund, A.-M. Henschen-Dahlquist, L. Lindahl, L. Nordenfelt, and J. Odelstad, editors, *Logical Theory and Semantic Analysis*, volume 63 of *Synthese Library*, pages 1–14. Springer, Netherlands, 1974.

[Lewis, 1981] D. Lewis. Ordering semantics and premise semantics for counterfactuals. *Journal of Philosophical Logic*, 10:217–234, 1981.

[Lindström, 1991] S. Lindström. A semantical approach to nonmonotonic reasoning, 1991. Uppsala Prints and Preprints in Philosophy, Department of Philosophy, University of Uppsala.

[Loewer and Belzer, 1983] B. Loewer and M. Belzer. Dyadic deontic detachment. *Synthese*, 54:295–318, 1983.

[Makinson, 1989] D. Makinson. General theory of cumulative inference. In M. Reinfrank, J. de Kleer, M. Ginsberg, and E. Sandewall, editors, *Proceedings of the 2nd International Workshop on Non-monotonic Reasoning*, volume 346 of *Lecture Notes in Computer Science*, pages 1–18. Springer, New York, 1989.

[Makinson, 1993] D. Makinson. Five faces of minimality. *Studia Logica*, 52(3): 339–379, 1993.

[McNamara, 1995] P. McNamara. The confinement problem: How to terminate your mom with her trust. *Analysis*, 55(4):310–313, 1995.

[Moulin, 1985] H. Moulin. Choice functions over a finite set: A summary. *Social Choice and Welfare*, 2(2):147–160, 1985.

[Nipkow *et al.*, 2002] T. Nipkow, L.C. Paulson, and M. Wenzel. *Isabelle/HOL: A Proof Assistant for Higher-Order Logic*, volume 2283 of *LNCS*. Springer, Lecture Notes in Computer Science, 2002.

[Nute, 1997] D. Nute, editor. *Defeasible Deontic Logic*. Kluwer Academic Publishers, Dordrecht, 1997.

[Olivetti and Pozzato, 2005] N. Olivetti and G. Pozzato. KLMLean 1.0: a theorem prover for logics of default reasoning. In H. Schlingloff, editor, *Proceedings of the 4th International Workshop on Methods for Modalities (M4M-4)*, pages 235–245. 2005.

[Olivetti and Pozzato, 2014] N. Olivetti and G. Pozzato. NESCOND: an implementation of nested sequent calculi for conditional logics. In S. Demri, D. Kapur, and C. Weidenbach, editors, *Automated Reasoning - 7th International Joint Conference, IJCAR 2014*, volume 8562 of *Lecture Notes in Computer Science*, pages 511–518. Springer, 2014.

[Parent, 2001] X. Parent. Cumulativity, identity and time in deontic logic. *Fundam. Inform.*, 48(2-3):237–252, 2001.

[Parent, 2008] X. Parent. On the strong completeness of Åqvist's dyadic deontic logic G. In R. van der Meyden and L. van der Torre, editors, *Deontic Logic in Computer Science (DEON 2008)*, volume 5076 of *Lecture Notes in Artificial Intelligence*, pages 189–202. Springer, Berlin/Heidelberg, 2008.

[Parent, 2010] X. Parent. A complete axiom set for Hansson's deontic logic DSDL2. *Logic Journal of the IGPL*, 18(3):422–429, 2010.

[Parent, 2012] X. Parent. Why be afraid of identity? In A. Artikis, R. Craven, N. K. Cicekli, B. Sadighi, and K. Stathis, editors, *Logic Programs, Norms and Action–Essays in Honor of Marek J. Sergot on the Occasion of His 60th Birthday*, volume 7360 of *Lecture Notes in Artificial Intelligence*, pages 295–307, Heidelberg, 2012. Springer.

[Parent, 2014] X. Parent. Maximality vs. optimality in dyadic deontic logic. *Journal of Philosophical Logic*, 43(6):1101–1128, 2014.

[Parent, 2015] X. Parent. Completeness of Åqvist's systems E and F. *Review of Symbolic Logic*, 8(1):164–177, 2015.

[Prakken and Sergot, 1997] H. Prakken and M. Sergot. Dyadic deontic logic and contrary-to-duty obligations. In Nute [1997], pages 223–262.

[Rott, 2001] H. Rott. *Change, Choice and Inference*. Clarendon Press, Oxford, 2001.

[Schlechta, 1995] K. Schlechta. Preferential choice representation theorems for branching time structures. *Journal of Logic and Computation*, 5(6):783–800, 1995.

[Schlechta, 1997] K. Schlechta. *Nonmonotonic Logics*. Springer, 1997.

[Sen, 1969] A. Sen. Quasi-transitivity, rational choice and collective decisions. *The Review of Economic Studies*, 36(3):381–393, 1969.

[Sen, 1971] A. Sen. Choice functions and revealed preferences. *The Review of Economic Studies*, 38(3):307–317, 1971.

[Sen, 1997] A. Sen. Maximization and the act of choice. *Econometrica*, 65(4): 745–779, 1997.

[Shoham, 1988] Y. Shoham. *Reasoning About Change: Time and Causation from the Standpoint of Artificial Intelligence*. MIT Press, Cambridge, MA, USA, 1988.

[Spohn, 1975] W. Spohn. An analysis of Hansson's dyadic deontic logic. *Journal of Philosophical Logic*, 4(2):237–252, 1975.

[Stalnaker, 1968] R. Stalnaker. A theory of conditionals. In N. Rescher, editor, *Studies in Logical Theory*, pages 98–112. Blackwell, Oxford, 1968.

[Steen and Benzmüller, 2018] A. Steen and C. Benzmüller. The higher-order prover Leo-III. In D. Galmiche, S. Schulz, and R. Sebastiani, editors, *Automated Reasoning–9th International Joint Conference, IJCAR 2018, Proceedings*, pages 108–116. Springer, 2018.

[Steen, 2018] A. Steen. *Extensional Paramodulation for Higher-Order Logic and its Effective Implementation Leo-III*, volume 345 of *DISKI*. Akademische Verlagsgesellschaft AKA GmbH, Berlin, 9 2018. Dissertation, Freie Universität Berlin, Germany.

[Stolpe, 2020] A. Stolpe. Unsettling preferential semantics. *Journal of Philosophical Logic*, 49:371–399, 2020.

[Temkin, 1987] L. S. Temkin. Intransitivity and the mere addition paradox. *Philosophy and Public Affairs*, 16(2):138–187, 1987.

[Tomberlin, 1981] J.E. Tomberlin. Contrary-to-duty imperatives and conditional obligation. *Noûs*, pages 357–375, 1981.

[van Benthem, 2001] J. van Benthem. Correspondence theory. In D. Gabbay and F. Guenthner, editors, *Handbook of Philosophical Logic*, volume 3, pages 325–408. Kluwer, Dordrecht, 2nd edition, 2001. Originally published in [Gabbay and Guenthner, 1984, pp. 167–248].

[van der Torre and Tan, 1997] L. van der Torre and Y.-H. Tan. The many faces of defeasibility in defeasible deontic logic. In Nute [1997], pages 79–121.

[van der Torre and Tan, 1999] L. van der Torre and Y. H. Tan. Contrary-to-duty reasoning with preference-based dyadic obligations. *Annals of Mathematics and Artificial Intelligence*, 27(1-4):49–78, 1999.

[van Fraassen, 1972] B. C. van Fraassen. The logic of conditional obligation. *Journal of Philosophical Logic*, 1(3/4):417–438, 1972.

[van Kutschera, 1974] F. van Kutschera. Normative Präferenzen und bedingte Gebote. In H. Lenk, editor, *Normenlogik*, pages 137–165. Verlag Dokumentation, Pullach bei Munchen, 1974.

[Veltman, 1985] F. Veltman. *Logics for Conditionals*. PhD thesis, University of Amsterdam, 1985.

Appendix A: Proof of Thm 3.3 (vi)

It is enough to describe a selection function model $M = (W, \mathfrak{f}, v)$ in which $\mathfrak{f}$ meets syntax-independence ($\mathfrak{f}$0), inclusion ($\mathfrak{f}$1), Chernoff ($\mathfrak{f}$2), consistency-preservation ($\mathfrak{f}$3) and Aizerman ($\mathfrak{f}$4), and in which (Sp) is falsified. The claim that (Sp) is not derivable in $\mathbf{F}$+(CM) follows at once from Theorem 4.13 (iii). The same holds for (RM).

Our counter-model for (Sp) is similar to the model used in the proof of Fact 2.13. Define $M = (W, \mathfrak{f}, v)$ as follows: $W = \{a, b, c\}$; $\mathfrak{f}$ is defined

by

$$\mathfrak{f}(A) = \begin{cases} \{a, c\} & \text{if } \|A\| = W \\ \|A\| & \text{otherwise;} \end{cases}$$

$v(p) = W$, $v(q) = \{b, c\}$, $v(r) = \{a, c\}$ and $v(s) = \emptyset$ for all the other propositional atoms s. (f0), (f1), (f2), (f3) and (f4) hold. But (Sp) is falsified at, e.g., world a:

- $\mathfrak{f}(p) = \{a, c\} \cap \|q\| = \{b, c\} \neq \emptyset \Rightarrow a \models P(q/p)$
- $\mathfrak{f}(p) = \{a, c\} \subseteq \|q \rightarrow r\| = \{a, c\} \Rightarrow a \models \bigcirc(q \rightarrow r/p)$
- $\mathfrak{f}(p \wedge q) = \{b, c\} \not\subseteq \|r\| = \{a, c\} \Rightarrow a \not\models \bigcirc(r/p \wedge q)$

Appendix B: Proof of Thms 4.2 (ii) and 4.5 (ii)

For the reader's convenience, I restate the theorems to be proven:

> *Theorem 4.2 (ii).* Under the max rule, **F**+(CM) is sound and complete with respect to the class of preference models in which $\succeq$ is max-smooth and transitive.
>
> *Theorem 4.5 (ii).* Under the max rule, **F**+(CM) is sound and complete with respect to the class of preference models in which $\succeq$ is max-smooth, transitive, and reflexive.

Soundness is straightforward. Completeness for models in which $\succeq$ is max-smooth and transitive follows from completeness for models in which $\succeq$ is max-smooth, transitive and reflexive. Therefore, I will focus on the latter. I find it more convenient to use an indirect approach, and show how the result can be obtained from the completeness theorem for a betterness relation max-smooth and reflexive, Theorem 4.5 (i), page 32. The detailed proof of the latter result may be found in [Parent, 2014]. The betterness relation in the canonical model as defined there does not satisfy the property of transitivity. Nevertheless, the desired result follows, because one can transform the model into one in which $\succeq$ is transitive in a truth-preserving way.[21]

Call $\succeq$ virtually connected whenever $a \succeq b$ implies $a \succeq c$ or $c \succeq b$. Given reflexivity, virtual connectivity implies totalness, but not the other way around. In [Parent, 2014] it is argued that on the canonical model

[21] A direct proof is also possible. We need only change the definition of $\succeq$ in the canonical model, and adapt the initial proof accordingly. The definition used by Goble for his systems DDL-4 [Goble, 2015, p. 176 *et seq.*] and DDL-c [Goble, 2019] achieves the result we want. The definition puts $(a, B) \succeq (b, C)$ whenever $(a, B) = (b, C)$ or $(B \geq C$ and $C \notin a)$. For simplicity's sake, I choose the indirect method.

of **F**+(CM) as defined there (cf. Definitions 4.11 and 4.12, page 37 *supra*) the betterness relation $\succeq$ is total (hence reflexive) and opt-smooth (*resp.*, max-smooth). The first step is to realize that $\succeq$ is also virtually connected, because the relation $\geq$ (in terms of which $\succeq$ is defined) is transitive. Recall that $A \geq B$ is a shorthand for $\bigcirc(A/A \vee B)$, and that (in the principal case) $(a, B) \succeq (b, C)$ iff: either $C \not\geq B$ or $B \in b$.

The following fact from [Parent, 2014] will also be helpful:

Fact B.1. *If $A \geq B \geq C$, $w^A \subseteq a$, and $C \in a$, then $w^B \subseteq a$.*

Proof. This is [Parent, 2014, Lemma 2 (iii)]. □

Now for the main observation:

Fact B.2. *In the canonical model $M^{(w,A)}$ of **F**+(CM) (as defined in Definitions 4.11 and 4.12, on page 37), $\succeq$ is virtually connected.*

Proof. Let (a, B), (b, C) and (c, D) be such that $(a, B) \not\succeq (c, D)$ and $(c, D) \not\succeq (b, C)$.

Case 1: $w^A \subseteq w$ for some A. In that case, the canonical model generated by (w, A) is as in Definition 4.11. So $C \geq D$, $D \geq B$ and $D \notin b$. From the first two, $C \geq B$, by law ($\geq$-trans) in Theorem 3.3. By construction, $w^C \subseteq b$. By (Id), $D \in w^D$ and so $w^D \not\subseteq b$. By Fact B.1, $B \notin b$. By Definition 4.11 (ii), $(a, B) \not\succeq (b, C)$ as required.

Case 2: $w^A \subseteq w$ for no A. In that case, the canonical model generated by (w, A) is as in Definition 4.12. When it is supposed that $(c, D) \not\succeq (b, C)$, that entails that $(b, C) \in \widetilde{W}$, by definition of $\succeq$. Either (i) $(a, B) := (w, A)$ or (ii) $(a, B) \in \widetilde{W}$. In case (i), $(a, B) \not\succeq (b, C)$ as required. In case (ii), the hypothesis $(a, B) \not\succeq (c, D)$ entails that $(c, D) \in \widetilde{W}$, and the claim follows for the same reason as in case 1. □

The second step is to realize that in the presence of reflexivity virtual connectivity and transitivity do not make much difference as long as we are only interested in maximal elements. To be more precise, a reflexive and virtually connected relation can be transformed into a reflexive and transitive (albeit not necessarily total) one in a truth-preserving way with respect to the max rule. (It does not matter which rule is applied in the input model, since its betterness relation is total.)

Theorem B.3. *For every preference model $M = (W, \succeq, v)$ in which $\succeq$ is reflexive and virtually connected, there is a preference model $M' = (W, \succeq', v)$ (with W and v the same) in which $\succeq'$ is reflexive and transitive, such that M and M' are equivalent under the max rule. Furthermore, if $\succeq$ is max-smooth, then $\succeq'$ is max-smooth.*

Proof. Starting with $M = (W, \succeq, v)$, define $M' = (W, \succeq', v)$ by putting $a \succeq' b$ whenever $a = b$ or $b \not\succeq a$.

Reflexivity of $\succeq'$ is immediate. Transitivity of $\succeq'$ follows from virtual connectivity of $\succeq$. Let $a \succeq' b$ and $b \succeq' c$. If one of $a = b$, $b = c$ and $a = c$ is the case, then we are done. So assume $a \neq b$, $b \neq c$ and $a \neq c$. Then $a \succeq' b$ and $b \succeq' c$ mean that $b \not\succeq a$ and $c \not\succeq b$. By virtual connectivity, $c \not\succeq a$, and so $a \succeq' c$ as required.

To show equivalence, it is enough to note that:

Lemma B.4. $\succ = \succ'$.

Proof of Lemma B.4. The argument for the $\subseteq$-direction appeals to the reflexivity of $\succeq$. Let $a \succ b$. Hence $a \succeq b$ but $b \not\succeq a$. The latter implies $a \succeq' b$, but also that $a \neq b$ (since $\succeq$ is reflexive). On the other hand, $a \succeq b$ and $a \neq b$ in turn imply $b \not\succeq' a$. Hence $a \succ' b$ as required.

For the $\supseteq$-direction, let $a \succ' b$. Hence $a \succeq' b$ but $b \not\succeq' a$. The latter means that $a \neq b$ and $a \succeq b$. For $a \succeq' b$ to hold, it must be the case that $b \not\succeq a$, which suffices for $a \succ b$. □

With Lemma B.4 in hand, the argument is straightforward since we have that, under the inductive hypothesis,

$$\max_{\succeq}(\|B\|^M) = \max_{\succeq'}(\|B\|^{M'}) \tag{2}$$

It is also straightforward to show that max-smoothness of $\succeq$ implies max-smoothness of $\succeq'$. Details are omitted. □

From this, Theorem 4.5 (ii) follows quickly. Suppose $\Gamma \nvdash_{\mathbf{F}+(\mathrm{CM})} A$. A similar argument as in the proof of Theorem 5 of [Parent, 2014] yields that the universe of the canonical model M of $\mathbf{F}$+(CM) contains a point a such that under the max rule a verifies all of Γ and falsifies A. On that model $\succeq$ is reflexive, max-smooth and virtually connected, Fact B.2. By Theorem B.3, M can be transformed into a model M' whose relation $\succeq'$ is reflexive, transitive and max-smooth. The two models share the same universe, so a is in M'. Under the max rule a verifies all of Γ and falsifies A, since the two models are equivalent. Thus, it is not the case that under the max rule $\Gamma \models A$ over the class of models in which the betterness relation is reflexive, transitive and max-smooth.

Appendix C: Proof of Thms 4.7 and 4.8

For the reader's convenience, I restate the theorems to be proven:

> *Theorem 4.7.* Under the max rule, **E** is sound and complete with respect to (i) the class of models in which $\succeq$ is transitive, and (ii) the class of models in which $\succeq$ is transitive and reflexive.
>
> *Theorem 4.8.* Under the max rule, **F** is sound and complete with respect to (i) the class of models in which $\succeq$ is transitive and max-limited, and (i) the class of models in which $\succeq$ is transitive, max-limited and reflexive.

Soundness is straightforward. For the completeness half, it suffices to invoke the following theorem.[22]

Theorem C.1 (Goble [2015; 2019])**.** *For every model $M = (W, \succeq, v)$, there is a model $M' = (W', \succeq', v')$ in which $\succeq'$ is reflexive and transitive, and such that under the max rule M and M' are equivalent. Furthermore, if $\succeq$ is max-limited, then $\succeq'$ is also max-limited.*

Proof. Let $M = (W, \succeq, v)$. Define $M' = (W', \succeq', v')$ as follows:

- $W' = \{\langle a, b, n\rangle \mid a, b \in W, n \in \omega\}$
- $\langle a, b, n\rangle \succeq' \langle c, d, m\rangle$ iff (1) $\langle a, b, n\rangle = \langle c, d, m\rangle$ or

$$(2)\begin{cases} \text{(a)}\ b = d\ \&\ n \geq m \\ \text{and} \\ (\text{b}_1)\ c \neq d\ \&\ a = c \text{ or } (\text{b}_2)\ c = d\ \&\ a \succ c \end{cases}$$

- $v'(p) = \{\langle a, b, n\rangle \mid a \in v(p)\}$

The following applies.

Fact C.2. *$W' \neq \emptyset$.*

Proof. This follows from the fact that $W \neq \emptyset$. □

Fact C.3. *$\succeq'$ is reflexive.*

Proof. This follows at once from clause (1) of the definition of $\succeq'$. □

Fact C.4. *$\succeq'$ is transitive.*

[22] [Goble, 2019, p. 44] describes the theorem as a modification and generalization of a theorem due to myself, planned for inclusion in the current chapter. At the time Goble wrote his chapter, such an inclusion was indeed planned. But Goble's result leaves out certain non-essential details, and for this reason I have decided to include it instead.

Proof. Assume $\langle a, b, n\rangle \succeq' \langle c, d, m\rangle$ and $\langle c, d, m\rangle \succeq' \langle e, f, l\rangle$.

In case one of these holds by clause (1) of the definition of $\succeq'$, then we are done. So suppose both hold by clause (2). By $(2.a)$, we have $b = d$ and $d = e$, from which $b = e$ follows. We also have $n \geq m$ and $m \geq l$. By transitivity of $\geq$, one gets $n \geq l$.

Note that $\langle a, b, n\rangle \succeq' \langle c, d, m\rangle$ and $\langle c, d, m\rangle \succeq' \langle e, f, l\rangle$ cannot hold in virtue of $(2.b_2)$ and $(2.b_1)$, respectively. The first implies $c = d$, while the second implies $e \neq f$ and $c = e$. One then gets $e = c = d = f$, a contradiction. Similarly, $\langle a, b, n\rangle \succeq' \langle c, d, m\rangle$ and $\langle c, d, m\rangle \succeq' \langle e, f, l\rangle$ cannot both hold in virtue of $(2.b_2)$. For in this case, $c \succ e$ and $e = f = d$ would imply $c \succ d$, and so $c \succ c$, given that $c = d$. This contradicts the irreflexivity of $\succ$. I consider the remaining cases in turn.

Suppose $\langle a, b, n\rangle \succeq' \langle c, d, m\rangle$ and $\langle c, d, m\rangle \succeq' \langle e, f, l\rangle$ both hold in virtue of $(2.b_1)$. In that case, $c \neq d$, $a = c$, $e \neq f$ and $c = e$. From $a = c$ and $c = e$, one gets $a = e$, and so we are done.

Suppose $\langle a, b, n\rangle \succeq' \langle c, d, m\rangle$ holds in virtue of $(2.b_1)$ and $\langle c, d, m\rangle \succeq' \langle e, f, l\rangle$ holds in virtue of $(2.b_2)$. In that case, $c \neq d$, $a = c$, $e = f$ and $c \succ e$. One gets $a \succ e$, and so we are done.[23] □

Lemma C.5. *Under the max rule, M' is equivalent to M. That is, for all $a, b \in W$, and all $n \in \omega$, $a \models A \Leftrightarrow \langle a, b, n\rangle \models A$.*

Proof. By induction on A. I only handle the case where $A = \bigcirc(C/B)$. For the left-to-right direction, it will help to note that, under the inductive hypothesis,

Sub-lemma C.6. *If $\langle c, d, m\rangle \in \max_{\succeq'}(\|B\|^{M'})$, then $c = d$.*

Proof of Sub-lemma C.6. Assume that $\langle c, d, m\rangle \in \max_{\succeq'}(\|B\|^{M'})$ and that $c \neq d$. We have $\langle c, d, m\rangle \models B$. Also $\langle c, d, m + 1\rangle \in W'$. By the inductive hypothesis, $\langle c, d, m + 1\rangle \models B$. Since $c \neq d$, we have

$$\langle c, d, m + 1\rangle \succeq' \langle c, d, m\rangle$$

But $m + 1 > m$, and so

$$\langle c, d, m\rangle \not\succeq' \langle c, d, m + 1\rangle$$

Thus, $\langle c, d, m\rangle \notin \max_{\succeq'}(\|B\|^{M'})$, contrary to assumption, and one must conclude that $c = d$, after all. □

[23] Fact C.4 is Lemma 31 in [Goble, 2019, p. 33]. I have modified the part of the argument dealing with the case where the two opening suppositions hold in virtue of $(2.b_2)$. In the paper the case is described as a possible one. But it is not, because the second supposition would hold only if (in the author's notation) $c < e$; this is a contradiction since $c = e$.

One can now turn to the proof of equivalence, starting with the right-to-left direction.

($\Leftarrow$) Assume $\langle a, b, n\rangle \models \bigcirc(C/B)$. Let $c \in \max_{\succeq}(\|B\|^M)$. We have $c \models B$. By construction $\langle c, c, n\rangle \in W'$. Assume for a reductio that $\langle c, c, n\rangle \notin \max_{\succeq'}(\|B\|^{M'})$. By the inductive hypothesis, $\langle c, c, n\rangle \models B$. So there is some $\langle d, e, m\rangle \in \|B\|^{M'}$ such that

$$\langle d, e, m\rangle \succeq' \langle c, c, n\rangle \qquad (\alpha)$$

$$\langle c, c, n\rangle \not\succeq' \langle d, e, m\rangle \qquad (\beta)$$

By (β), $\langle c, c, n\rangle \neq \langle d, e, m\rangle$. Thus, (α) holds because condition (2.a) of the definition of $\succeq'$ is met along with one of $(2.b_1)$ and $(2.b_2)$. Since $c = c$, $(2.b_2)$ applies, viz. $d \succ c$. By the inductive hypothesis, $d \models B$. But, then, $c \notin \max_{\succeq}(\|B\|^M)$. So one must conclude that $\langle c, c, n\rangle \in \max_{\succeq'}(\|B\|^{M'})$. But one then gets $\langle c, c, n\rangle \models C$ from the opening assumption. By the inductive hypothesis, we get $c \models C$, which suffices for $a \models \bigcirc(C/B)$.

($\Rightarrow$) Assume $a \models \bigcirc(C/B)$. Let $\langle c, d, m\rangle \in \max_{\succeq'}(\|B\|^{M'})$. By the inductive hypothesis, $c \models B$. By Sub-lemma C.6, $c = d$, *viz.* $\langle c, d, m\rangle$ is $\langle c, c, m\rangle$. Assume for a reductio that $c \notin \max_{\succeq}(\|B\|^M)$. There is some d such that $d \models B$ and $d \succ c$. But $\langle d, c, m+1\rangle \in W'$. By the inductive hypothesis, $\langle d, c, m+1\rangle \models B$. By the definition of $\succeq'$, one gets $\langle d, c, m+1\rangle \succ' \langle c, c, m\rangle$, a contradiction. So one must conclude that $c \in \max_{\succeq}(\|B\|^M)$. From the opening assumption, $c \models C$, and so $\langle c, d, m\rangle \models C$ by the inductive hypothesis. This establishes the desired claim $\langle a, b, n\rangle \models \bigcirc(C/B)$. □

It remains to verify that, if $\succeq$ is max-limited, then $\succeq'$ is max-limited. Assume that there exists some $\langle a, b, n\rangle \in W'$ such that $\langle a, b, n\rangle \models A$. By Lemma C.5, $a \models A$. Since $\succeq$ is max-limited, there is some c with $c \in \max_{\succeq}(\|B\|^M)$. Re-running the same argument as that for the right-to-left-direction of Lemma C.5, one gets $\langle c, c, n\rangle \in \max_{\succeq'}(\|A\|^{M'})$, and thus $\succeq'$ is max-limited.

Xavier Parent
Institute of Logic and Computation, Technical University of Vienna, Vienna, Austria.
Email: xavier@logic.at

2

Recent Thought on *Is* and *Ought*: Connections, Confluences and Rediscoveries

Lloyd Humberstone

Abstract. Section 1 of this critical survey recalls the much discussed difficulty noted by A. N. Prior in a 1960 paper for a formulation of Hume's Law to the effect that no valid inference can take us from non-ethical premises to an ethical conclusion. Section 2 presents a response by Toomas Karmo from the 1980s, echoes of which surface in discussions of the problem over the past 5–10 years, also noted in this section along with some objections that have been raised to this line of thought. Section 3 reviews another, more recent (2010) contribution to the debate, from Greg Restall and Gillian Russell, and discusses its connections to the material in play in previous section, as well as aspects of the reception of this contribution by commentators. This way of organizing things makes possible a reasonably comprehensive guide to (at least the main highlights of) the recent literature. Several more detailed passages are postponed to Postscripts at the end of each section, or demoted to footnote discussion, to be skipped by those wanting a speedier overview, though of necessity that will mean that some voices go unheard and some mistakes uncorrected.

1 Introduction **72**

2 Karmo Recalled **89**

3 The Restall–Russell Approach **119**

I am grateful to Raphael Morris for numerous suggestions and corrections as this paper was being composed, and to John Horty for his subsequent assistance.

Handbook of Deontic Logic and Normative Systems, Volume 2

1 Introduction

What might reasonably count for present purposes as *recent* in the literature on the principle variously called Hume's Law, the Is–Ought Gap, or the thesis of the autonomy of ethics, is perhaps given by the publication date – 2010 – of the stimulating, varied, and much discussed anthology Pigden [2010], with perhaps special mention due to the paper Restall and Russell [2010] therein, in view of its ambitious elegance and the interest it has sparked in subsequent forays into the field. On a slightly larger time scale, recency might be dated back to 1988 and the appearance of the strikingly original Karmo [1988], to which little attention gets paid in Pigden [2010].[1] We should be alive to the possibility these and other alternative responses to, in particular, difficulties for Hume's Law raised in Prior [1960], are not, once terminological adjustments are made, mutually incompatible, and they accordingly compete for our attention rather than for our assent. The proponent of one such response wants to focus on one aspect of the situation while those favouring an alternative reaction are essentially saying, "No, let's look at things *this* way." The present discussion is not entirely neutral, expressing a particular interest in the Karmo-style approach, but with an even greater interest in looking at some of links that emerge between various responses to Prior, touching also, if sometimes all too briefly, on several post-[Pigden, 2010] discussions (in chronological order of publication: Brown [2014] and [2015], Singer [2015], Wolf [2015], Maguire [2015], Woods and Maguire [2017] and Fine [2018]). Slightly less recent contributions, some before and some since Prior [1960], will also be touched on. After the present introduction, Section 2 looks at the content, subsequent discussion, and sometimes unknowing re-discovery, of aspects of Karmo 1988, though this theme also finds its way into a final Section 3, similarly focused on Restall and Russell [2010] and its reception.

We need, therefore, to begin by recalling the nub of Prior 1960. Suppose E and F are respectively uncontroversially an ethical and a non-ethical statement, in the latter case supposing – even if one thinks that this does not hold automatically in virtue of the classification of F as non-ethical – that $\neg F$ is also non-ethical. One might have wanted a version of Hume's Law saying that no ethical statement is a logical conse-

[1] It is, however, mentioned in Maitzen's contribution to the collection, [Maitzen, 2010]. [Maitzen, 1998] paid it much closer – albeit unsympathetic – attention, dialectically downstream from which we have [Nelson, 2007; Hill, 2008; Maitzen, 2008; Hill, 2009].

quence of a consistent set of non-ethical statements, where "non-ethical" just means "not ethical".[2] We ask about the status of the disjunction $E \vee F$, and observe that since this follows from F it must be classified as non-ethical to avoid a violation of the envisaged law, but then from the nonethical $E \vee F$ and $\neg F$, there follows our ethical conclusion E, giving a different violation of Hume's Law. So there is no way to classify $E \vee F$ which permits us to retain Hume's Law.[3]

Prior's discussion features several further examples with a less artificial flavour to them than the disjunctive example just abstractly rehearsed, including several about what all undertakers ought to do or what should be done to all New Zealanders, which raise some distractions it would be helpful to be able to avoid. Before doing so, let us note that even these examples, which will be familiar to anyone who has dipped into the Prior-initiated dialectic on all this, and rather artificial. For this reason, Jackson [2013][4] offers something closer to a real life example:

> Suppose Jane has serious reservations about abortion but nevertheless agrees to pay for a close friend's abortion. For her the *a priori* valid inference
>
> > I have paid for an abortion.
> > Therefore, if anyone who has paid for an abortion has done something morally wrong, I have done something morally wrong.

[2]One would not normally have to say this, but I see that, after using 'non-ethical' for many pages, in note 12 on p. 59 of [Brown, 2014], Brown casually remarks: "As I use the term 'non-ethical', it is not equivalent to 'not ethical'. To say that a sentence is neither ethical nor non-ethical is, therefore, no violation of the Law of Excluded Middle."

[3]Rescher [1990], note 2, mistakenly says that this argument was first given in [Mavrodes, 1968]. Mavrodes, apparently not familiar with Prior [1960], gives the argument and does indeed provide an excellent discussion of the issues it raises, with many deft moves, several appearing here in notes 6, 14 and 33. The snapshot of Prior's argument given above conceals some details brought out in the proof of Proposition 1.1 in the Postscript to this section. Forty years after Prior's paper, Sinnott-Armstrong gives the same argument in [Sinnott-Armstrong, 2000] (and again in a mild re-working of this material as Chapter 7 of [Sinnott-Armstrong, 2006]) with no mention of either Prior or Mavrodes. (Among other things, the re-working corrects a typo from line 5 of p. 171: "$2 + 2 = 5$" to "$2 + 2 \neq 5$".) No doubt Prior's argument has occurred independently to many people; the present author thought he had discovered it in the late 1970s and was lucky enough to have a better-informed colleague (Edward Khamara) who directed him to Prior's discussion.

[4]This is one of three entries in the encyclopedia in which it appears, all of them directly addressing, in their own ways, the issue under discussion here; the other two are [Elgin, 2013] and [Pigden, 2013].

corresponds to a line of reasoning that worries her a great deal.

Now, even if this example seems straightforward, we should recall that the nature of conditionality in deontic contexts has been notoriously problematic. Conditionals with "ought" apparently in their consequents force us to decide between constructions – using "O" as the *Ought* or *Obligation* operator of deontic logic – of the form $p \to Oq$, $O(p \to q)$ and $O(q/p)$. In the first two cases here, $\to$ is our default representation for material implication, though it could be swapped out for another (e.g., subjunctive) conditional construction, and the third features the primitive binary conditional obligation operator $O(\cdot/\cdot)$ of dyadic deontic logic, itself open to competing semantic interpretations,[5] and with status of Modus Ponens for whatever format is adopted being the subject of perennial debate.[6] When we pass to universal generalizations of the "Whatever is F ought to be G", complications ramify further: $O\forall x(Fx \to Gx)$, $\forall xO(Fx \to Gx)$, $\forall x(Fx \to O(Gx))$, $\forall xO(Gx/Fx)$,...[7] An instructive example of the issues arising from trying to assign the appropriate scope to "O" when formally representing some of these constructions with *if*, *all* and *ought* is given in the report at p. 10 of Mares [Mares, 1992] on a spat with a referee for that paper. After some involvement with these concepts in the following paragraph, we will accordingly do our best to steer clear of them.

[5]References to many alternative semantic proposals for this connective are listed under Example 4.4.4, p. 241, in [Humberstone, 2016].

[6]A tiny sample, in chronological order: Greenspan [1975], Humberstone [1983], Section 7.4 of Makinson [1999], Kolodny and MacFarlane [2010], Saint Croix and Thomason [2014]. The 'fundamental problem' of the title of Makinson's paper concerns the problematic status of truth-based semantics for those who think of normative language as not truth-valued, rather than the specifics of conditional constructions. A good first move in the solution of that problem is made (on p. 58*f.*) of [Mavrodes, 1968]: if that's how you feel about truth, just run the discussion in terms of an artificial predicate stipulated to behave disquotationally. In terms of this, let's say, schmuth-predicate, we have: "People ought never to lie" is schmue if and only if people ought never to lie. (Mavrodes actually uses 'right' rather than 'schmue', but this introduces distractions. A point similar to Mavrodes' is made in note 4 of [Singer, 2015]: "one may substitute whatever analogue of truth one wishes here.") This is a first step because in the semantics for deontic logic we need not just the absolute notion of (something like) truth, but a world-relativized notion – preferably still 'thin' enough as not to beg any questions against non-cognitivism.

[7]One may be tempted to include on this list "$O\forall x(Gx/Fx)$", thinking that it may be true that in respect of each of the Fs it would be better that it be G than not ($\forall xO(Gx/Fx)$, on one common understanding), and at the same that it would be disastrous if all Fs were G. But the envisaged addition to the list is not well formed, since dyadic O takes two formulas to make a formula: the slash separates these two, rather than being part of a restricted quantifier notation.

Rynin [1957] had considered arguments apparently of the form 'Ga, Therefore $O(Ha)$' which might be felt to be enthymematic and, with the missing premise $\forall x(Gx \to O(Hx))$ restored, would no longer be (at least blatant) counterexamples to Hume's Law. Rynin then cleverly executes a conditional proof step[8] (not that he describes it in exactly these terms) to pass from the now explicit form:

$$Ga, \forall x(Gx \to O(Hx)) \vdash O(Ha) \tag{1}$$

to

$$Ga \vdash \forall x(Gx \to O(Hx)) \to O(Ha) \tag{2}$$

In concrete terms, Rynin writes [1957, 314*f.*]: "Thus if we have the argument: 'I have given my promise, and all promises ought to be kept, therefore I ought to keep my promise', we can transform it into 'I have given my promise, therefore if all promises ought to be kept, then I ought to keep my promise.' " The simplified version (2) represents "a is a promise; therefore, if all promises ought to be kept then a ought to be kept."[9] We see already with this example that the point about scope arises: should the 'all promises ought to be kept' have been $O\forall x(Gx \to Hx)$ instead, in principle undermining the validity of the pre-conditional-proof version of the argument.[10]

In fact, rather than discussing such $\vdash$-claims as (1) and (2) here, Rynin [1957, p. 314] discusses what he calls the conditionals corresponding to the arguments thereby represented, where the main conditional is reproduced here (using "$\dashv\!\exists$") as strict implication,[11] and using (possibly

[8] This move is also made in [Pigden, 2016]: see (B♯) on p. 407.

[9] In Section 2 we will encounter Searle's idea in [Searle, 1969] that the it may be possible to drop this premise altogether from the original argument, because of an analytic connection between having made and not yet kept a promise, on the one hand, and being such that one ought to keep it, on the other. Indeed, Rynin is already sympathetic to such a view, speaking (p. 316) of a "normative principle that serves as a rule of inference to validate the derivation" – though this is not a part of Rynin's discussion that Prior takes up.

[10] Wolf [2015] raises justified doubts about the example 'Lois should donate to charity if she is able' and the unobviousness of whether this has the form *Lois is able to donate to charity* $\to O($*Lois donates to charity*$)$, on the one hand, or $O($*Lois is able to donate to charity* $\to$ *Lois donates to charity*$)$, on the other. What may be less justified is the association of this example with p. 264*f.* of Vranas, where the closest case resembling this one concerns instead the sentence 'If Jane is a citizen, she ought to vote,' especially as Vranas is maintaining that something about these examples – their genuine normativity, if not their logical form – varies from case to case.

[11] In fact Rynin writes "e→" here for an entailment connective clunkily defined as the conjunction of a strict implication with a conjunct saying that its antecedent and

decorated) "N" and "F" to represent normative and factual statement represents the transition from (1) to (2) as a transition from the one conditional to the other, i.e., as from

$$(N \wedge F) \nrightarrow N' \qquad \text{to} \qquad F \nrightarrow (N \to N').$$

This representation is potentially problematic if the N and F here are taken as playing the E and F roles above, since, the N is already what is in the literature (and below) called a *mixed* case, the main operator not being O, which instead governs only the consequent of a (universally quantified) material conditional here.

Prior follows Rynin with variations on (2). Since all that the work the universally premise is doing here is done by the single instance with a as x, however, we might as well just simplify, both (1) and (2) by rewriting the universal premise to $Ga \to Ha$, which turns (2) into

$$Ga \vdash (Ga \to O(Ha)) \to O(Ha). \tag{3}$$

But if $\vdash$ here is taken as the consequence relation of classical logic, the right-hand side is equivalent to $O(Ha) \vee Ga$, so we are again considering essentially the $E \vee F$ case of the second paragraph above. It is not being suggested that this was Prior's own route from the rather cluttered natural Rynin-style examples to the streamlined – though less natural-seeming – disjunction case.[12] However, since, as already remarked, it is far from clear how to handle natural language deontic conditionals, it is safer to avoid the issue as much as possible, and stick to uncontroversially Boolean embeddings and interactions with our monadic deontic

consequent are not analytic. This second conjunct (for which one might have expected – the equally clunky – "and neither the consequent nor the negation of the antecedent is analytic") can be ignored for present purposes, though. Rynin's dot notation for conjunction has also been replaced here with "$\wedge$". In Section 2 we will be discussing an approach to these matters according to which strict implication, understood as truth-preservation in all worlds, is a good conceptualization of entailment, which instead has to be taken as truth-preservation relative to all worlds and all ethical (or normative) standards.

[12] Prior explicitly thanks [1960, p. 202] one T. H. Mott for suggesting it to him, and was in any case at around the time at which [Prior, 1960] was written, unaware of the classical equivalence of $p \vee q$ and $(p \to q) \to q$ (or $(q \to p) \to p$), as we see from the 'Note 1960' appended (p. 229) to the discussion of deontic logic in [Prior, 1962], retracting his recent favourable remarks about $Op \to \big((p \to Oq) \to Oq\big)$ as a plausible deontic principle: no-one realising that this was another way of writing $Op \to (p \vee Oq)$ would find it at all plausible, especially with the further reformulation – again recalling that the logical background here is classical – to $(Op \to p) \vee Oq$: either all obligations are fulfilled or everything is obligatory. (The present observation is adapted from p. 476, last ten lines, in [Humberstone, 1995].)

operators. The case of $\vee$-introduction is especially simple in illustrating the presence of material in the conclusion of a valid argument not present in the premises, undermining, as Rynin [1957] pointed out, an attempt by P. H. Nowell-Smith in an attempt to establish Hume's Law as a special case of the supposed impossibility what is thereby illustrated. Rynin and Prior diagnose this as a case of overfamiliarity with syllogistic reasoning at the expense of the fuller picture provided by (then) contemporary logic. Pigden [2016, p. 403], turns up an appeal to this same incorrect principle from a 1725 publication – though that is more understandable, since the syllogism was then the only game in town.[13]

Reactions to Prior's disjunctive syllogism argument naturally include whose querying the underlying logic assumed in delivering the claimed consequences – the $\vee$ introduction step in passing from F to $E \vee F$ (queried in [Beall, 2014]) or the disjunctive syllogism step taking us from $E \vee F$ and $\neg F$ to E (queried in [Mares, 2010]); the Postscript to this section begins by taking up the second of these reactions, which urges as a remedy for this unfortunate malady: a shift from classical to relevant logic.[14] The first response will be touched on at the end of Postscript (i)

[13]The relevant considerations do appear to take some time to absorb. Garcia [1995, p. 549], reconstructing Hume's reasoning in the famous is–ought passage offers as a version of one of its premises: "No proposition with an 'ought'-operator governing some element within it can be deduced from a group of propositions none of which contains this feature." Garcia's comment on this is that it comes close to assuming the desired conclusion to begin with, rather than that, taken at face value (and with 'proposition' corrected – so that it makes better sense – to something more linguistic, such as 'sentence' or 'statement') it is simply false.

[14]Mares observes that the ingredients in Prior's argument – $\vee$-introduction (or "addition" as some of those in our bibliography say) along with disjunctive syllogism – are those involved in the C. I. Lewis/Albert of Saxony demonstration [Anderson and Belnap, 1975, p. 164] that any contradiction has every statement as a (classical) consequence. But care is required with this observation – the care displayed by Mavrodes [1968] as he considers what he lists as Objection 5 to his/Prior's argument. (In Mavrodes' presentation of the argument our F, E, become F, M, respectively, and specific but representative choices are made as to which statements these are. F is 'The Fisher Building is the tallest building in Detroit', and M is 'Men ought never to lie'. Their disjunction is called D.) The objection says that the argument trades on the controversial feature of classical logic that a mutally contradictory statements together entail everything, calling only for a revision of Hume's Law to exclude inconsistent premises. Mavrodes (p. 362) then writes concerning this objection, that "... in the form given here it is simply mistake about the structure argument which I have discussed. I have nowhere used or discussed any argument which includes both F and not-F (or any other contradiction) among its premises. I have instead pointed out that if D is normative then it is entailed by F, and hence there is a nonnormative statement which entails a normative one. On the other hand, if D is nonnormative, then D and not-F together entail M, which again subverts the gap thesis. Now

to Section 3. Turning to responses not contesting the underlying logic, which is standardly taken to be classical logic (though for the inferences mentioned so far – not including the point about the implicational definability of disjunction, of course – could equally well be intuitionistic logic), we have what we might call *trichotomy* responses. These retain the emphasis on a failure of anything in some class – call it the 'conclusion class' – to follow from a set of statements in another class – call it the 'premise class'. Here, uncontroversially (or 'basic') ethical statements are in the conclusion class, while the similarly straightforward nonethical statements are in the premise class; but these two classes are not jointly exhaustive of all the statements.[15] For instance the premise

neither of these entailments involves any self-contradictory premises. One of them has only the single premise F, and the other has the pair of premises D and not-F. But neither of them involves the contradictory premises F and not-F."

[15]This strategy is dubbed the "No Mixed Sentences Defence" by Campbell Brown and discussed by him in Section 3, bearing that title, of [Brown, 2014]. At least, so it seems at the start of that section. As we proceed, however, it transpires that Brown doesn't mean by "mixed" what is usually meant by this: that we have some basic ethical statements and some basic non-ethical statements, and the mixed cases arise as combinations of the one with the other using Boolean connectives and quantifiers. (This is what "mixed" has meant in these discussion for over fifty years, occurring with this signification in [Atkinson, 1958] and [Schurz, 2010] from 1958 and 2010 respectively, and of course in many other contributions in between.) By the time we get to p. 58, however, we are worrying about sentences which contain, on the one hand no ethical, and on the other, no non-ethical predicates – as though being 'mixed' amounted to having both ethical and non-ethical *vocabulary*. (Here, for Proposition 1, there is an appeal to an implicational formulation of the Halldén completeness of first-order predicate logic without identity, to show that there are no implications from formulas without ethical predicates to formulas in which only ethical predicates appear, a corollary of Prop. 1 called 'NOFI 3' – No Ought from Is, Mark 3 – by Brown. This corresponds to the Barrier Lemma on p. 472 of Humberstone [1982a], for a propositional logic with two sets of sentence letters, one set for the basic ethical and the other for the basic non-ethical case – but the latter was not envisaged to represent statements constructed with *only* ethical and logical vocabulary.) Sentences entirely devoid of non-ethical vocabulary are surely of negligible interest from the perspective of Hume's Law, and are certainly not basic ethical sentences. (But see also note 65 below.) By contrast with basic ethical statements, for Brown, "[p]urely ethical sentences are rarely encountered in the wild, outside the philosopher's laboratory. Notice, for example, that even Prior's sentence 'All New Zealanders ought to be shot' fails to be wholly ethical (assuming 'New Zealander' is non-ethical). Two oddities with this comment: first, there is no 'even' about it – on the second page of [Prior, 1960], we have: "I would not count as 'ethical' a statement in which only ethical and logical expressions occurred essentially." Secondly, why is only 'New Zealander' mentioned and not also '(are) shot' as non-ethical vocabulary – albeit non-ethical vocabulary embedded in the scope of a deontic operator making the whole of "ought to be shot" an ethical – though not what Brown calls a 'purely ethical' – expression?

class might be described as *factual*, the conclusion class as *ethical*,[16] and the rest *mixed*. The contrasting *dichotomy* response[17] aims at a formulation of Hume's Law in which every statement gets to be in either the premise class or the conclusion class. That seems closer to the letter of Hume's own formulation about conclusions containing 'ought' being claimed to be derivable on the basis of premises not containing 'ought', even if such an overly syntactic characterization would be a hopeless first stab at a precise articulation of the spirit of Hume's discussion. Still, one would like some exhaustive non-ethical/ethical division with a significant inferential relation that can be seen never to take us from the former to the latter. Differently put, a dichotomous approach aims at a claim of closure: the class of non-ethical statements is closed under something like entailment. Prior's $\vee$-introduction + disjunctive syllogism argument shows that any once-and-for-all way of redistributing the slack the mixed cases comprise into the one of the two classes to obtain such a dichotomy approach cannot succeed, when that inferential

[16]Or perhaps *evaluative* or *normative*, though these terms are generally understood to encompass much more than the specifically ethical or moral. These broader notions create a potential problem of their own for the present discussion of the validity of arguments with such-and-such premises and so-and-so conclusions, if they treat *valid* itself as an evaluative terms, as does Urmson [1953, p. 223]: "to call an argument valid is not merely to classify it logically, as when we say it is a syllogism or *modus ponens*; it is at least in part to evaluate or appraise it; it is to signify approval of it." See also, in this connection, Shaw [1965], where considerable play is made of seemingly valid arguments *about* arguments which conclude with verdicts on the latter arguments' validity or invalidity, despite not having any of their premises evaluative. One might try to abstract from any such evaluativity, saying that for logical purposes validity is to be understood as no more than the necessary, *a priori*, or formally secured (depending on the purposes at hand) preservation of truth, but even truth itself has been held to be an evaluative or normative notion: see Horwich [2018] for a discussion of several thinkers (which do not include Horwich himself) inclined to say such things. Certainly at some point along the line from 'Snow is white' through 'The proposition that snow is white is true' to 'The belief that snow is white is correct', we seem to have gone from the non-normative to the normative. This calls for comment even if the normativity is not ethical: the puzzle is formulated and addressed in [Gibbard, 2005].

[17]The dichotomy/trichotomy terminology for marking this contrast appears in [Schurz, 1997] and [Schurz, 2010]. A dichotomous version of Hume's Law is called the Special Hume thesis (or 'SH') in these publications (and in [Schurz, 1994]), in which Schurz looks at conditions on bimodal alethic–deontic logics necessary and sufficient for them to satisfy SH. The simpler monomodal version of such results appears as Lemma 5.6 in [Zolin, 2000]; further characterizations of the class of logics concerned, called (fully) modalized logics, can be found in [Humberstone, 2016, §4.6]. Potentially confusingly, Morscher [2016] uses the term *dichotomy* for the (as [Morscher, 2016] puts it) descriptive/normative contrast even when summarizing Schurz's trichotomous SH findings.

relation is taken to be entailment; for a more precise statement, see the Postscript. We should be alive the possibility that the way that redistribution is effected may need to influence the replacement of entailment proper – even when the latter is subject to the further requirement that the premises are consistent ('closure under consistent consequence' as it is put in the Postscript). A more promising candidate will emerge in Section 2.

This issue of how to distribute the slack is in large part a technical problem rather than one of special meta-ethical significance, the latter more aptly applying to the unmixed cases in the trichotomous approach: the basic ethical and basic non-ethical cases. Even the significance of the latter (non-exhaustive) division was subjected to serious questioning by Peter Singer in [1973], where the serious gap is taken to be that between recognising that things are thus-and-so in the world on the one hand and taking this as a reason for acting a certain way, on the other. It is not so important whether what is considered one's moral beliefs are taken as tied to the former recognition or to the latter acknowledgment – assimilations associated in the 1960s with Philippa Foot and R. M. Hare respectively, and called non-neutralism and neutralism by Singer.[18] The focus in what follows is mostly on moral language of the deontic rather than the axiological kind, and even here one faces a choice as to whether to concentrate, for example, on unnegated – or more generally unembedded – *ought*-judgments, or to include also negated such judgments, often spoken of (in the distinctive English associated with deontic logic) in terms of permissibility.[19] The latter will be the policy here: a claim

[18]Singer cites Hare explicitly though not Foot, but p. 52, left column, lines 5–6, makes it clear he has Foot in mind by illustrating it with a principle about *clasping one's hands* three times an hour as, according to the neutralist, a candidate moral principle, held as such a principle by those ordering their lives by resolutely acting in accordance with it. The same point was later made in [Jaggar, 1974], esp. Section V. Neutralism about the content of morality is evidently more congenial to of an internalist inclination, wanting to minimize the step from moral judgment to disposition to act.

[19] The standardly quoted passage from Hume's Treatise – to be found in many of the entries in our bibliography – is open to a respectable interpretation as specifically focusing on unnegated (etc.) *ought*-conclusions, and, for instance, the opening page of [Mares, 2010] takes Hume's Law specifically to pertain to the underivability of formulas of the form OA from sets of formulas free of deontic vocabulary. (However, in mid-p. 123, Wolf [2015] cites a case from the Treatise in which Hume queries an inference from premises about human nature to a permissibility conclusion – i.e., to a negated *ought* judgment.) This includes the cases in which A itself contains further deontic vocabulary, excluded under the rubric 'single-main-occurring O-conclusion' below. See also Mares [1992], where it is shown for a relevant deontic logic favoured there that for deontic-free A, B, the implication $A \to OB$ is never provable: see what

that such and such is, for example, *not* morally required is just as much a moral claim as the claim that whatever is under consideration *is* morally required.[20] Those favouring a normal deontic logic will not need to as a further as a case to consider that of conjunctions of *ought*-judgments, in view of the equivalence in such a logic of $OA \wedge OB$ with $O(A \wedge B)$.[21]

Another case worthy of consideration is the disjunction of two *ought*-statements, as Daniel Singer (2015, p. 196) reminds us, saying of a proposal from Gibbard [2012] to consider only straight unembedded occurrences of "O":[22] "It is too strong because it excludes some arguments from the purview of the is-ought gap that it should not. For instance, it excludes an argument with the conclusion 'Either Jane ought to eat tomato soup, or Ange ought to buy garlic bread.' " Here, an argument with this conclusion normal (indeed, more generally, monotone) deontic logics would have $O(A \vee B)$ as a consequence of $OA \vee OB$ despite the failure of the converse implication so typically one could still conclude to a single-main-occurring O-conclusion, but (*i*) this might end us up with a conclusion that few would consider ethical despite the main O (e.g., if B is $\neg A$) and (*ii*) this would not work with agent-relative or agent-implicating deontic operators (as arguably in the example of Jane and

Mares calls Lemma 2.5 (though it is not actually used to prove anything else) on p. 14; on the other hand, the logic in question does contain theorems of the form $A \to \neg OB$ for deontic-free A, B, such as $\Box\neg p \to \neg Op$ – contraposing an observation from the base of p. 15. Thus here it is only the 'main O in the conclusion' form of Hume's Law that holds (and indeed the single-main-O form that is being shown to hold).

[20]Space considerations preclude a discussion of the question of moral nihilism here, which this cursory remark raises – a topic arising in several of the publications referred to; in particular: [Maitzen, 1998; Maitzen, 2010; Hill, 2008; Nelson, 1995; Pigden, 2007], and the final section (headed §8.6) of Maguire [2015]; also relevant is the discussion of 'positively ethical' sentences in the §5 of Brown [2014]. (There is a typo in the first new paragraph of p. 68 here, with "any sentence implied by an inconsistent sentence is inconsistent", presumably intended to read "any sentence implying an inconsistent sentence is inconsistent.") Maitzen (2010, p. 307*f.*) regards the Disjunctive Syllogism part of Prior's argument, the transition from $E \vee F$ and $\neg F$ to E as straightforwardly refuting Hume's Law since each of the premises, but not the conclusion, is ethical by the following criterion: each is capable of being accepted by a moral nihilist (which doesn't mean, we may take it, that both could be simultaneously accepted by such a nihilist.) Sinnott-Armstrong [2000, p. 161 second paragraph], endorses a similar principle.

[21]Early opposition to this equivalence can be found in [Schotch and Jennings, 1981]; for other references, see the index entries under 'aggregation' in Humberstone [2016], which refers specifically to the implication from $OA \wedge OB$ with $O(A \wedge B)$, though even the converse implication has been contested – e.g. in Jackson [1985].

[22]Maitzen [2010] recalls with approval a broadly similar suggestion from Gewirth [1979].

Ange) – though for simplicity we ignore such operators in what follows.[23] Finally, let us the case of conditionals – which to avoid the complications alluded to above – we may take to involve material implication of the form $OA \to OB$. We saw these emerge, above, from the 'conditional proof' move made in Rynin's dialectic from [Rynin, 1957], and we can also see them in play in [Sen, 1966], which is of some interest in having prompted Hare to write (replying to Geach [1976a]) the following [Hare, 1977, p. 469]:

> I have indeed been persuaded, not by Geach but by Professor Amartya Sen, that my own thesis of universalizability commits me to allowing valid inferences from non-evaluative premises to logically complex evaluative conclusions.

Sen [1966, esp. p. 76] is mostly concerned with good rather than ought, and with inferences from "A and B are descriptively alike" to "A being good implies B being good", though the latter can be reformulated with a slight change of meaning so the that conclusion is instead "A is as good as B," making though less readily dismissible as a Boolean compound of evaluative sentences by anyone not considering such cases to fall within the basic ethical category (not that specifically moral goodness is at issue in the cases discussed by Sen, who also presents similar examples involving *ought*). Here we are in the vicinity of the issue of the supervenience of the ethical on the non-ethical, whose connection to Hume's Law is a much discussed matter, the discussion using requiring a consideration of contrasts between metaphysical and logical (or more broadly, conceptual) necessity that it is accordingly preferable to avoid here.[24]

Here we take the 'generous' line that all of these Boolean compounds are candidates for being 'basic ethical', even if some (as in the $OA \vee \neg OA$ case no less than the non-embedded $O(A \vee \neg A)$ case mentioned above) warrant exclusion, a topic to which we return in the Postscript to this section, after Proposition 1.1 there. (Recall that the Boolean compounds at issue here do not include the problematic 'mixed' compounds.) In particular, the case of negated *ought*-judgments this has the effect that any

[23] For references to the extensive literature on them, see [Humberstone, 2016], p. 251. As to the "single" in "single-main-occurring", the intention is to set aside encoding, for example $\neg OA$ or $OA \vee OB$ as of the desired simple form by rewriting them as $O\neg OA$ or $O(OA \vee OB)$ (or even $O(OA \vee B)$) to which they would be equivalent in the logic KD45, for instance. For more details and qualifications concerning Gibbard on the present issue, see Singer's discussion, including note 9 on p. 196 of [Singer, 2015].

[24] An airing of some of the relevant considerations and a look at the main literature can be found in Section 8 of [Humberstone, 2019].

one-premise inference from an *ought*-premise to a non-ethical conclusion will contrapose to an inference, valid if and only the original is, from a non-ethical premise to a basic ethical conclusion, as is often remarked in the case of the 'ought'-implies-'can' principle, contraposing to such things as 'Sylvie is unable to attend her mother's funeral' to 'it is permissible for Sylvie not to attend her mother's funeral'. The point is hinted at on p. 313 of [Rynin, 1957], where Rynin suggests that "[i]n fact, most people hold many views similar in nature so far as entailment of factual by normative or normative by factual statements is concerned. In saying this I do not mean to assert that most people use the word 'entails' or have ever heard it used, but that they would agree, say, that no one is under any obligation to do what he cannot do." Though Rynin had earlier (p. 309) noted the general point about contraposition, it is more explicitly brought to bear with ought-implies-can in the contraposition point is more was made more explicitly in [Mavrodes, 1964].[25] A minimal pertinent observation would be that while contraposing the conclusion of an argument with one of its premises preserves validity, it does not preserve the property of being a potentially explanatory argument, or the property of recording a justification for accepting the conclusion on the basis of the premises (cf. note 33 below); a good discussion of these issues is provided by Basl and Coons [2017].

Section Postscript: Logical Considerations Arising. Apropos of the emphasis on disjunctive syllogism in Mares – and indeed the title of – [Mares, 2010], we should recall, in addition to the remarks from Mavrodes quoted in note 14, the following.[26] If the class of ethical statements, or indeed any class of statements whatever, is deemed to be closed under taking negations and under converse entailment, then it can be shown to contain all statements if it contains any, by means of a chain of reasoning appearing in diagrammatic form as Figure 3 on

[25] A (comparatively) recent discussion of the Ought-implies-Can thesis with Hume's Law very much in mind can be found in [Vranas, 2007], which also provides an extensive survey of the literature, including its pre-history (see note 3 there). Heading (2) under note 1 of [Vranas, 2007] lists in chronological order many who have suggested that the 'implies' in Ought-implies-Can should really be '(semantically) presupposes', in which case the contraposition step fails. The list begins with [Atkinson, 1958], to which we can add (from the following year) [Remnant, 1959]. The still more recent [Vranas, 2018] on Ought-implies-Can bears less closely on our current concerns.

[26] Maitzen (1998, note 11), also recognises the potential for an objection to such disjunctive syllogism moves on relevant-logical grounds but takes it that the particular use he wants to make of such a move will not be one that will raises relevantist objections.

p. 135 of [Humberstone, 1996], headed "A Lewis-like argument immune to relevance objections," rather than "A Prior-like argument immune to relevance objections". That is because of the connections (much emphasized in [Humberstone, 1996]) between the material under considerations here and the treatment of subject matters presented in [Lewis, 1988], touched on below in the Postscript to Section 3. This argument can also be found in note 5 (p. 192) of [Maguire, 2015], and on p. 153 of [Russell, 2010].[27] The caption reference to Lewis rather than Prior arises the argument here alluded to because if one thinks of the task at hand as that of moving from a trichotomous basic-ethical/basic-nonethical/remainder to a dichotomous ethical/nonethical division for a formulation of Hume's Law, then the assumption that the ethical category in the two-block division – subsuming now many mixed cases formerly housed in the 'remainder' category – will be closed both under negation and under converse entailment lacks the appeal that such a closure assumption might have for the initial 'basic' ethical class and factual classes. Indeed, we do not need even that assumption for the disjunctive syllogism case. We just need, to recall our $E \vee F$ case, *just a single* non-ethical statement F with a non-ethical negation, in order to pass from the ethical $F \vee E$ (so classified because if it were ethical, by the converse entailment condition – alias the one-premise version of Hume's Law – F some chosen non-ethical statement would be ethical after all) together with the *ex hypothesi* $\neg F$, to E, given the counterexample to Hume's Law in its two-premise form. No general 'closure under negation' principle is appealed to here, just the assumption that *some* basic non-ethical

[27] Russell remarks (p. 160, note 3) that she "came across this argument in Gideon Rosen's Spring 2001 graduate seminar at Princeton." Instead of trotting it out again here, I will give a variant. Suppose we have a non-empty class of statements closed under taking negations and under converse entailment. Let A be an element of this class, and B be an arbitrary statement, with a view to showing that B is also an element of the class and hence that from its non-emptiness it follows that the class contains all statements. By the negation condition $\neg A$ is in (the class) since A is. So by the converse entailment condition $\neg A \wedge \neg B$ is in; so by the negation condition $\neg(\neg A \wedge \neg B)$ is in; (a redundant step this next one, to make the reasoning easier to follow) so by the converse entailment condition $A \vee B$ is in; and so, finally, by the converse entailment condition again, B is in. All of this reasoning is fine in the system FDE of first-degree entailment, with $\neg$ taken as the favoured De Morgan negation, a common core of relevantly accepted principles before one even considers the addition of a relevant implicational connective to the language and what its logical properties might be: see [Anderson and Belnap, 1975, §15]. (As Guevara [2008] mentions, an argument along these lines, with the specific is–ought case in mind, appears already on p. 468 of [Humberstone, 1982a].)

statement has a non-ethical negation.[28] By contrast, it turns out, as we shall see in Proposition 3.10, that Prior style arguments make essential appeal to a two-premise rule (disjunctive syllogism or some substitute), whereas the 'linear' Lewis-like arguments from [Humberstone, 1996] and note 27 assume negating mixed conjunctions and disjunctions keeps us on the same side of the extended ethical/nonethical divide, but appeals only to one-premise inference rules.

This last point was insufficiently emphasized in Humberstone [1996], especially as the re-worked version of the account in [Humberstone, 1982a] is there explicitly noted not to satisfy the general condition that the negation of anything ethical (in a world – since this is a world-relative taxonomy) is again ethical in that world.[29] In view of that and also in view of Theorem 2 – labelled 'Prior's Dilemma' – in the Formal Appendix to Fine [2018], which gives something like Prior's argument in the setting of an abstract theory of propositions rather than of sentences of a formal or natural language, and explicitly assumes that the classes of descriptive and normative propositions are each closed under negations, it would seem worthwhile here to show that such global assumptions are not needed for at least the version of the argument as it appears in Prior [1960]. Certain aspects of the argument that were left tacit in the summary given above – and indeed are left tacit in [Prior, 1960] – are made explicit, very much along the lines of Fine's discussion (except for the closure-under-negation assumption). It should be added also Mares is quite right to observe that *this* reasoning would not go

[28] Guevara [2008], p. 55, writes: "It is widely held that sentences containing 'ought,' or other normative terms, are closed under negation. But I show that this is questionable." But of course the class of sentences containing expressions on any list you care to care to come up with *is* closed under negation, because the negating the sentence leaves whatever vocabulary the original sentence contained still intact – at least for a large class of natural languages, of which English is one (with the exception of few expressions – the 'positive polarity' items). It turns out that what Guevara has in mind is that the class of normative sentences is not closed under negation, where containing 'ought' and the like is not sufficient for normativity. In pointing forward (p. 46) to the passage just quoted, Guevara remarks similarly that "the concept of guidance I press throughout also calls into doubt an assumption – widely held – that sentences containing 'ought' or other normative terms are closed under negation," meaning that an *ought*-judgment's ability to guide the agent to a specific action type is not inherited by the permissibility judgment which results from negating it.

[29] Nor is mentioned in the earlier Geach [1979] where (p. 229) Prior's $\vee$-Introduction + disjunctive syllogism argument is given but with the gratuitously strong assumption that the premise class is closed under negation. (In fact Geach assumes this about the conclusion class as well, treating the two classes symmetrically from the start and thereby disposing of what he calls the theory – or range of theories – of *logical islands*.)

through in relevant logic (say, putting $\vdash_{\mathsf{FDE}}$ – see note 27 – in place of $\vdash_{\mathsf{CL}}$ below).

Consider the following suppositions we might make concerning a class of statements $\mathbb{F}$:

(1a) $F \in \mathbb{F}$ and (1b) $\neg F \in \mathbb{F}$

and we have another statement E about whose membership in $\mathbb{F}$ we make no assumption, but we do suppose,

(2) E and F are logically independent according to the consequence relation $\vdash_{\mathsf{CL}}$ of classical propositional logic, in the sense that for no binary truth-function $\#$ – notation we use now for the associated (not necessarily primitive) connective – do we have $\vdash_{\mathsf{CL}} E \# F$.

(3) $\mathbb{F}$ is closed under 'consistent consequence' in the sense for any CL-consistent $\{A_1, \ldots, A_n\} \subseteq \mathbb{F}$, if $A_1, \ldots, A_n \vdash_{\mathsf{CL}} B$ then $B \in \mathbb{F}$. (The consistency condition can be taken to mean that $A_1, \ldots, A_n \nvdash_{\mathsf{CL}} C$ for some C, or equivalently, given that $A_1, \ldots, A_n \vdash_{\mathsf{CL}} B$, that we do not also have $A_1, \ldots, A_n \vdash_{\mathsf{CL}} \neg B$.)

The letters E and F are intended to recall ethical and factual (or non-ethical), as in the presentation of Prior's argument in the main body of this section, and condition (3) with its consistency rider is taken from Prior [1960] too – *pace* 'Objection 5' considered in Mavrodes [1968], mentioned in note 14. (2) packs a lot into it, since (considering $\#$ and the binary first projection and negated first projection functions and likewise for the second coordinate) it implies that

$$\nvdash_{\mathsf{CL}} F \text{ and } \nvdash_{\mathsf{CL}} \neg F\text{; and } \nvdash_{\mathsf{CL}} E \text{ and } \nvdash_{\mathsf{CL}} \neg E,$$

as well as ruling out essentially binary relations: $E \nvdash_{\mathsf{CL}} F$ etc. (taking $\#$ as $\rightarrow$).[30] Finally, although we presume available the logical apparatus of classical propositional logic, any extension of that logic (by quantifiers, modal – e.g., deontic – operators, or whatever), is fine, and for sentences $C_1, \ldots, C_n, C_{n+1}$ of some such richer language, "$C_1, \ldots, C_n \vdash_{\mathsf{CL}} C_{n+1}$" means that there is a substitution s and there are formulas of the language of classical propositional logic proper, $A_1, \ldots, A_n, A_{n+1}$ with $s(A_i) = C_i$ $(1 \leq i \leq n+1)$ and $A_1, \ldots, A_n \vdash_{\mathsf{CL}} A_{n+1}$.

Proposition 1.1. *From assumptions (1)–(3) above, it follows tha t* $E \in \mathbb{F}$.

[30] This criterion of logical independence is that employed in Lemmon [1965]; a discussion of how to adapt it to independence relative to non-classical logics can be found in Humberstone [2020].

Proof. Since $F \in \mathbb{F}$ by (1a) and $\{F\}$ is consistent (by (2)), by (3) we have $E \vee F \in \mathbb{F}$. Now $\{E \vee F, \neg F\}$ is also consistent, since otherwise $E \vee F \vdash_{\mathsf{CL}} F$ and so $E \vdash_{\mathsf{CL}} F$ violating assumption (2). Therefore, since $E \vee F, \neg F \vdash_{\mathsf{CL}} E$, and we have not only $E \vee F \in \mathbb{F}$ but also (by (1b)) $\neg F \in \mathbb{F}$, by (3) we have $E \in \mathbb{F}$. □

The consistent closure condition (3) above is formulated by Prior [1960, p. 201] in terms of excluding 'self-contradictory' premises in a putative counterexample to Hume's Law, because from such premises "one could deduce not only ethical conclusions but any conclusions whatever, trivially," which will again invoke suspicions that $\vdash_{\mathsf{CL}}$ is showing its weakness here, but here we raise the issue to observe that, unlike some (e.g., Fine [2018]) there is no corresponding exclusion on the 'conclusion' side of conclusions B for which $\vdash B$ (or $\vdash = \vdash_{\mathsf{CL}}$ or again, any desired extension thereof): (3) does not have a further condition that B is not such a formula, even though that too would have trivialized the claim that it is a consequence of any $\{A_1, \ldots, A_n\}$. The reason is that Prior is taking it that any such B is automatically on the 'non-ethical' side of the fence (in our $\mathbb{F}$, that is), as he indicates on the previous page of [Prior, 1960], with the classification as non-ethical of such things as "It either is or is not the case that I should fight for my country" in which the ethical vocabulary occurs inessentially ($Op \vee \neg Op$ being such a special case of $A \vee \neg A$ with O replaceable by any sentence operator). As was mentioned in note 15, Prior goes on to add that not only should a statement to be classed as ethical contain ethical expressions (such as 'O' or *ought*, in the intended sense) essentially, but it should not contain, logical vocabulary aside, *only* such expressions, as in 'It is obligatory that what is obligatory be done' – one of Prior's favourite deontic principles, schematically: $O(OA \to A)$, the subject of Example 1.2 below. – though this is cited along with other popular candidate deontic axioms, so it is not completely clear whether here we are trading on their status as logical truths (those B for which $\vdash B$, with $\vdash$ a favoured consequence relation) or on the constituent vocabulary point officially being illustrated. For that we would have needed some such example as "It is obligatory that what is obligatory *not* be done," "If anything is permissible it is obligatory," the converse of another example Prior gives here (a vernacularized form of the famous D-for-'Deontic' axiom). Whereas the vernacular versions of candidate principles of deontic logic are explicitly on Prior's list of statements in which the presence of moral language does not occasion classification as ethical, we need to cases in which such language is 'de-activated' by appearing within belief

and indirect speech contexts.[31] In one sense, the statement that Jane feared that she had done the wrong thing is ethical in content, namely in the sense that grasping its content requires the possession of ethical concepts. But this is not the notion of ethicality that is at issue with Hume's Law. Finally, let us put on the record an important point, due to Campbell Brown [2015]: the use of something like disjunctive syllogism is essential to refuting Hume's Law with the likes of Proposition 1.1, where "something like" disjunctive syllogism means: like it in respect of being an at least *two*-premise rule of inference. We return to this in Proposition 3.10 in Postscript (i) to Section 3. We conclude with some words on what was described above as a favourite deontic principle of Prior's.

Example 1.2. Concerning the schema $O(OA \to A)$, or more accurately the claim that all instances of this schema are true, Prior tells use at p. 229 of [Prior, 1962] it "was originally suggested to me by Mrs. J. F. Bennett (in 1953 or 1954) as an example of a synthetic *a priori* proposition." For reasons of space, it has not been possible to discuss the Gideon Rosen's now well-known *flurg* argument (which can be found in Russell [2010], Guevara [2008], Singer [2015]), but the following simplified variant of the definition of *flurg* is presented by Guevara [2008, p. 48]:

> We might just as well have coined the term 'blurg' to mean 'to do something one ought not to do in any actual circumstances." This yields another valid inference from 'is' to 'ought' (. . .): 'Jones is in some actual circumstances. Therefore, Jones ought not to blurg.' Here we derive, apparently, a kind of categorical imperative against blurging. This confirms our sense that there is something shady about the style of counterexample, and that the problem with it must lie at least in part in the arbitrariness of the stipulated terms.

Since the reference to actual circumstances is vacuous here, let us write 'a performs action x' as 'Dax', so that 'a blurgs' is in effect defined to mean $\exists x(Dax \wedge O\neg Dax)$. Thus to say that a ought not to blurg is to say: $O\neg\exists x(Dax \wedge O\neg Dax)$, or, with some processing, $O\forall x(Dax \to \neg O\neg Dax)$, or again, $O\forall x(O\neg Dax \to \neg Dax)$, and instantiating the $\forall x$

[31] The active/inactive terminology here is taken from p. 201 of [Schurz, 2010] in connection with what Schurz calls the Max Weber Thesis (the fortunes of which he charts through a range of deontic–doxastic logics in §7.1 of [Schurz, 1997]); of course, the classification of contexts which are de-activating – or as it is put in Humberstone [1997], 'protective' – needs careful attention: doxastic contexts, yes, epistemic contexts, no (and so on).

to b, say, we have $O(O\neg Dab \rightarrow \neg Dab)$: so Guevara's categorical imperative emerges as a particular case of Mrs Bennett's synthetic *a priori* principle – all very Kantian rather than evidently calamitous, so perhaps not the reductio Guevera was hoping for. (For more on this principle, as a candidate modal axiom, see the index entry "**U**" in Humberstone [Humberstone, 2016] – that being the label associated with this axiom by Lemmon and Scott.) Similar considerations are raised by what turned out to be a contentious example in Geach [1977, p. 474*f.*], that of Evan and Dewi Williams in which a crucial (though Geach says 'vacuous') premise is "Nobody ought to adopt the practice of *doing something he ought not to at least twice every day*." The example and the description of this premise as vacuous certainly seemed to puzzle Hurka and Borowski in [1980] and [1980], respectively; communication is then further hampered by an impatient reply, [Geach, 1982], on Geach's part, affecting bafflement at Borowski's (fairly standard) deontic notation and citing in its note 4 the title of Borowski [1980] alongside with the publication details of Borowski [1976]. ◁

2 Karmo Recalled

The proposal of Karmo [1988] is in what we might call the Shorter-inspired family of responses to the problem of extending the ethical/non-ethical taxonomy from the basic cases so as to subsume the mixed cases in such a way that we end up with everything falling in line with the basic ethical statements or with the basic non-ethical statements, though which side they fall into line with depends on which contingent facts obtain. So we end up with a world-relative taxonomy which, relative to any given world, is a two-block partition and is to that extent a dichotomy style approach, though one which is, as we shall see, world-variably dichotomous. Coupled with this, one backs off from attention to arguments with the world-invariant property of validity to those with the world-relative property of soundness.[32] Shorter [1961] does not actually put matters in these terms and writes of futility or uselessness

[32]Here a sound argument is a valid argument with true premises (and, therefore, a true conclusion). When the premises are contingent premises, this makes the soundness of a valid argument from them to a conclusion a contingent matter. This is not the only use of the term *sound* (as applied to arguments or inferences) one will find in the literature. For example, in Chapter 1 of Lemmon [1965] "sound" is used to mean "valid"; a related example from the same period would be Shaw [1965]. (In fact, Shaw uses "sound" as replacement for "valid" in case the application of latter term should be held to be an entirely non-evaluative matter: see note 16.)

rather than unsoundness, and others (perhaps Karmo, even) may find this attribution of the approach to him contentious, so we devote a footnote to its defence.[33] The attribution in question was originally made in [Humberstone, 1982a] where an earlier world-relative taxonomy, perhaps less satisfactory than Karmo's (for reasons given in note 58 below),

[33]Shorter writes [1961, p. 286*f.*], "In A [[an F to $F \vee E$ inference, where A1 is the F premise and A2 the disjunctive conclusion]] it is clear that a specific ethical duty can be derived from A2 [[the conclusion of the inference]] only if we know that the first half of the disjunction is in fact false. If it is false then we can derive the duty (...) If it is true, then A2 is of no help to us in deciding whether [[the duty in question exists]]. But if the first half of A2 is false, then A1 is false; and if A1 is false then the inference A lends no support to the conclusion A2." So there is no world in which both the $\vee$-introduction inference and the disjunctive syllogism inference are sound. (We know this *a priori*, but of course it will typically be an *a posteriori* matter where the unsoundness lies; for example in the concrete version discussed by Mavrodes – see note 14 – the $\vee$-Introduction inference from F to $F \vee M$ is certainly unsound, whether or not the disjunctive syllogism inference is also unsound: the tallest building in Detroit at the time Mavrodes was writing [Mavrodes, 1968] was the Penobscot Building, not the Fisher Building.) But knowledge and its absence are mentioned in as well as the mere truth of the premises, though in his discussion of a second example on p. 287, Shorter stresses the role of knowledge. In this connection, it is worth recalling something said by Mavrodes (p. 363*f.*) after he wields Prior's argument to show, as Prior had done, that Hume's Law (put in terms of entailment or logical consequence) is mistaken "I have not even attempted to establish the corresponding epistemological thesis, i.e., that we could come to know some normative statement on the basis of some nonnormative statement. Nor will I attempt to do so here." One explanation of this is that to acquire knowledge of one thing on the basis of knowledge of another the inference in question would need to be *sound* and not just valid – which is not to say that soundness would *suffice* in this connection: see the last sentence of (the main body of) Section 1 above. Sinnott-Armstrong [2000] also repeatedly raises the issue of the soundness as opposed to the mere validity of arguments violating Hume's Law (apparently unaware – see note 3 – of Prior, Shorter, Karmo or anyone other than Nelson [1995], who is similarly unaware of Karmo's earlier discussion of essentially his main argument), though again his chief concern is which the justificatory efficacy of such arguments, remarking at p. 167: "Thus, even if Hume's doctrine fails logically, if it works epistemologically, then that might be enough to serve the primary purposes of many defenders of the doctrine." Similarly Heathcote is apparently similarly unaware of the attempts to use this consideration to adjust the version of 'Hume's Law' facing Prior-style difficulties; not this undermines the content of what he says, writing [Heathcote, 2010, p. 94]: "[N]ote that Hume is concerned with what can be discovered through reasoning: thus his division is a division of *sound deductive inference*, not of merely valid deductive inference. Nowhere does Hume imply that his division corresponds to what we think of as valid deductive inference." In the present discussion, to avoid over-use of the term 'argument' since what Prior gives us is an argument (in part) about arguments, the term *inference* is used as a substitute for the 'inner' arguments or the associated argument forms ($\vee$-introduction and disjunctive syllogism), rather than to suggest that their conclusions might characteristically be arrived at by inference from their premises.

and about which Karmo makes some comparative remarks in note 7 of [Karmo, 1988]. (The passage in question is quoted at the start of the Postscript to this section.) The suggestion from [Humberstone, 1982a] is briefly recalled at the end of Postscript (ii) to Section 3 below.

There may even be a semi-conscious anticipation of the world-relative approach – or the rejection of taxonomic essentialism, as Maitzen [2010] calls it – in Prior [1960, p. 204]

> If a conclusion containing an expression E is validly inferred from a certain premise or set of premises, and the inference would remain valid if E were replaced by any expression whatever of the same grammatical type, then I say that in that inference the expression E is contingently vacuous. The expression "ought to" is in this sense contingently vacuous in the inferences "Tea-drinking is common in England, therefore either tea-drinking is common in England or all New Zealanders ought to be shot"(...)

Attention to the replaceability *salva validitate* of ethical vocabulary has been the focus of much subsequent work on Hume's Law – Jackson [1974], Pigden [1989] and [2010], Schurz [1994], Chapter 4 of [Schurz, 1997], and [Schurz, 2010], for example – but the point of current interest is not how precisely to formulate the relevant considerations or how they bear on apparent counterexamples (whether defusing them as objections or acknowledging them as counterexamples). The issue is, rather, Prior's choice of terminology: what is *contingent* about Prior's contingent vacuity? Of course it is contingent which expressions are of what Prior calls the same grammatical type, but this seems no more to warrant calling the occurrence of a token of such a type, relative to a given inference, 'contingently vacuous' than the fact that if the expression featuring in an inference has meant something different – as they might well have done – would warrant calling the inference 'contingently valid'. What is contingent here is the *truth* of the premise about tea-drinking, sufficing for the truth of any disjunction in which it is a disjunct, thereby nullifying the bearing of any ethical vocabulary in the remaining disjunct on the truth-value of the disjunction: the truth of the disjunction under these circumstances in no way hangs on how the application of that vocabulary. But had the first disjunct been false, everything would depend on how that vocabulary applied...

Similarly, Prior is hovering in the vicinity of a Shorter-style reaction when he writes (p. 201):

> Finally, in case my conditions are not stringent enough, I shall with all my examples proceed as follows: Wherever I claim that a

> certain statement is an ethical conclusion, and give a deduction of it from purely non-ethical premises, I shall also give a deduction of the same conclusion from premises which are not all non-ethical, and the deduction will be of a sort generally recognised as leading to an ethical conclusion. That is, to anyone tempted to query the "ethical" status of my conclusion, I shall say "Look, you can also get it *this* way"; and if that was where you had first met with it, you wouldn't have dreamed of denying its 'ethical' character".

But what is 'getting' the conclusion in this or that way? For his official position, this needs to be 'validly infer' – yet the persuasive effect of the examples could be due entirely to our understanding this as 'soundly infer': faced with the argument, we imagine that the premises are true and take it from there. The validity of the argument takes us overtly to the truth of the conclusion in the circumstances imagined, but perhaps more covertly to a particular verdict as to the ethicality of the conclusion in those circumstances.[34]

We need to hear from Karmo himself on all this. The references, in the following quotation, to what all parties to the debate would agree on calling ethical or agree on calling non-ethical may be taken as references to what we have been calling the basic ethical or non-ethical cases, respectively, and the examples alluded to were presented before this passage in [Karmo, 1988], two of them originating in Prior [1960]:

> To deal with such examples, we define a sentence S to be *ethical* in a possible world w just in case S is true in w with respect to one ethical standard, and false in w with respect to another ethical standard.
>
> We explain the term 'ethical standard' as follows. Call a sentence 'uncontroversially ethical' just in case all parties to the

[34]Similarly, Pigden, whose is–ought work has concentrated, like the others mentioned alongside him in the precedent paragraph, on replaceability *salva validitate* of ethical vocabulary in the conclusions of putative counterexamples exceptions to Hume's Law, quietly shifts the focus from validity to soundness at p. 221*f.* of [Pigden, 2010] in remarking that when we look at the conclusions on their own we agree that they may contain moral vocabulary essentially ("in a certain sense" – which I take to be the sense that they are not logically or *a priori* equivalent to sentences lacking the vocabulary in question) but under certain conditions such an equivalence does hold with arbitrary same-category replacements: "namely," Pigden writes, "when the premises of the arguments are true." (In fact, with the concrete examples, Pigden substitutes the predicate "hedgehog" for the moral vocabulary, as in his [Pigden, 1989], in order to underline the fact that a purely general logical point is involved here.) Pigden is picking up on the discussion at pp. 202–203 of Schurz [2010] in which (in)essentiality figures only in an argument-relative way and there is no move from validity to soundness.

> logical-autonomy-of-ethics debate would unite in calling it ethical. (There surely are sentences of this kind, for example, 'It ought to be the case that all New Zealanders are shot.' 'Everything that Alfie says is true' and 'Either tea-drinking is common in England or it ought to be the case that all New Zealanders are shot', on the other hand, are presumably not sentences of this kind: for agreement is presumably lacking on their status.) Then the ethical standard subscribed to by a person is completely determined once it is determined what truth values he assigns to all uncontroversially ethical sentences.
>
> We take it that any possible world can be uniquely picked out with some assignment of truth values to those sentences which the parties to the logical-autonomy-of-ethics debate would unite in calling non-ethical. We take it that just as some one possible world is the actual world, so some one ethical standard is the correct ethical standard. When people simply say, 'Sentence S is true', we take them to mean 'S is true in the actual world with respect to the correct ethical standard'. When people simply say, 'S is true in world w', we take them to mean 'S is true in w with respect to the correct ethical standard'.[35]

In a footnote (note 6) appended to this passage, Karmo suggests that for heuristic purposes we might think of the ethical standard as given by a set of ideal or perfect worlds in a simplified Kripke model for deontic logic, or more generally, one might add, the accessibility relation of a such a model.[36] What ought to be the case is what is the case in all the ideal (more generally, in all the accessible) worlds. Such models can be thought of simply as triples $\langle W, X, V\rangle$ with $X \subseteq W$ in the simplified case (or with X replaced by $R \subseteq W \times W$ in the general case), and V assigning appropriate semantic values to the non-logical vocabulary,[37] The reader is assumed to be comfortable with the inductively defined notion of the truth of a formula A at a point $w \in W$ in such a model, notated (for approximate conformity with Restall and Russell in [2010]) by writing

[35] Karmo [1988, p. 254]; I have added italics to 'ethical' in the first paragraph since this is where the term is being defined, and also italicized the world variable "w".

[36] We can make the simplification to a subset containing the ideal worlds when any two worlds have he same worlds accessible to them, in which case that common set will serve as the set of ideal worlds in one of the simplified – or as it is put in [Humberstone, 2016], *semi*-simplified – Kripke models. If this set is required to be non-empty, then the deontic logic determined by the collection of such models is that known as KD45.

[37] For example, in the case of propositional logic, V would map each sentence letter (or propositional variable) to a subset of W, an outright stipulation as to which worlds it is true at.

$\mathfrak{M} \models_w A$, where $\mathfrak{M}$ is, say, $\langle W, X, V\rangle$ and $w \in W$.[38] X would of course be replaced by $R \subseteq W \times W$ for the general case in with the simplification is not wanted. For Karmo's purposes the simplified version is very much what is wanted, though, because it secures the desired independence of the ethical standard and the non-moral facts taken to distinguish one world from another.[39]

Definition 2.1. *For any formula A, any model $\mathfrak{M} = \langle W, X, V\rangle$ and any $w \in W$, A is ethical at w in $\mathfrak{M}$ if and only if for some $X' \subseteq W$ with $\mathfrak{M}' = \langle W, X', V\rangle$, exactly one of the following is the case: $\mathfrak{M} \models_w A$, $\mathfrak{M}' \models_w A$.*

This definition of ethicality adapts the informal characterization given in the opening sentence of the passage quoted from Karmo above. (Following Karmo, when we are not explicitly relativizing to a model, we say "ethical in world such-and-such, but to avoid doubling the "in", when that relativization is in force, we say "ethical at such-and-such

[38] In fact Restall and Russell omit the valuation component V of the models, with the result that what are supposed to be models look more like *frames*, though since their discussion is in terms of truth rather than just validity they must be somehow thinking of the elements of a model as carrying with them the kind of semantic properties normally regarded as conferred on them by V. (Many others avoid a model component like V, which is specifically there to make semantic assignments to atomic expression, and instead incorporate in its place the satisfaction relation $\models$ itself, or some equivalent, such as $\| \cdot \|$, assigning semantic values to all expressions, including formulas/sentences. But Restall and Russell include no such device, though on pp. 21 and 253, at one point they use the notation "$w \Vdash p$" without making it clear how this is supposed to be construed, given their official notion of a model. Another option, often followed in computer science and AI-related applications of Kripke semantics, is to think of the points in a model as sets of sentence letters, or the associated characteristic functions, to start with. But whatever one thinks of the merits of this for alethic and deontic interpretations of modal logic, for the common tense-logical interpretation in which the points are moments of time, it leaves no room for the idea that two moments, one strictly later than the other, might verify precisely the same atomic sentences; Section 5.3 of [Restall and Russell, 2010] appeals to essentially this interpretation. It is for this extra flexibility that, when Scott [1974] explains the transition from matrix methodology to model-theoretic semantics using indexed bivalent valuations, it is the indices, not the valuations, that play the role of points a model.) Also, [Restall and Russell, 2010] uses, not the present models, which underlie the model-theoretic version of Karmo's discussion, but *pointed* models (and the above reference to frames should really be to pointed frames): we return to this in Section 3.

[39] 'Desired' here means: required for Karmo's project. As we shall see below, in discussing Daniel Singer's independent rediscovery of this way of handling matters, Woods and Maguire [2017] are highly critical of building in such an independence at this fundamental level, wanting an account that would leave open potentially contested meta-ethical perspectives.

world in so-and-so model".) A more direct adaptation would put after the 'if and only if' the following:

> for some $X', X'' \subseteq W$ with $\mathfrak{M}' = \langle W, X', V\rangle$, $\mathfrak{M}'' = \langle W, X'', V\rangle$ and exactly one of the following is the case: $\mathfrak{M}' \models_w A$, $\mathfrak{M}'' \models_w A$.

But this is equivalent to ethicality as per Definition 2.1 since given the latter we get this variant by taking $\mathfrak{M}''$ as $\mathfrak{M}$ and given the variant we get the original back by noting that if $\mathfrak{M}'$ and $\mathfrak{M}''$ differ in respect of verifying A at w, one of them must agree in that respect with $\mathfrak{M}$'s treatment of A at w. Non-ethicality at w in $\mathfrak{M}$ is of course just the negation of this, and so amounts to a formula's having the same truth value at w however X – our current simple-minded incarnation of the model's ethical standard – is varied.[40] The informal use made in Section 1 of talk of basic ethical and basic non-ethical statements can be understood as represented here, for a given model, as meaning ethicality at all worlds in the model and ethicality at none of them, respectively.[41]

Karmo's own characterization of world-relative ethicality should be taken as the analogue of Definition 2.1 for natural language declarative sentences in place of formulas of a formal language, with respect to something playing the role of an intended model. The more formalized version is presented here to aid comparison in the following section with the similarly model-theoretic discussion in Restall and Russell [2010].

[40] Karmo [1988], at the end of note 6 there, mentions the richer option of using instead a *betterness* relation on the worlds as playing the ethical standard role, in order to handle conditional obligation statements, and yet further variations would need to be incorporated to handle not only deontic but axiological vocabulary ('morally good' etc.), where the standard would specify the application-conditions for the predicates concerned in terms of non-moral features of the individuals or actions they apply to. But here we are concerned with the fundamental ideas of Karmo's picture and how they bear on the debate over Hume's Law (which was itself similarly formulated by Hume in deontic terms – *ought* and *ought not*).

[41] Admittedly this may not sit well with Karmo's gloss 'uncontroversially ethical', since such things as "James should visit his mother in hospital" can be understood as uncontroversially ethical – deemed ethical by all parties to the is–ought debate, that is – though obviously not true in all ideals worlds, in some if not all of which James' mother is not in hospital to begin with. It would perhaps be better to speak of fundamental moral principles rather than uncontroversially ethical statements, in this case; the 'M-class' as opposed to 'm-class' statements of Basl and Coons [2017] would be another contender (to the extent that it differs from the basic principles/derived judgments distinction). A fully developed version of Karmo's position would need to address this matter more thoroughly than the rather sketchy treatment in [Karmo, 1988] does.

And, as just mentioned, the only reference to the sets of ideal or permissible worlds in Karmo's suggestion in his note 6, as a simple concrete realization of the concept of an ethical standard, the main discussion being cast in the latter terminology, somewhat abstractly conceived. We stick with the concrete suggestion here, in part so that it the concepts in play can be clearly illustrated in Examples 2.2. For these illustrations we concentrate on a simple deontic incarnation of the schematically presented $E \vee F$ case from the second paragraph of Section 1. F was to be 'basic' non-ethical, so we take it as a sentence letter p, and E, basic ethical, so let it be Oq (q another sentence letter, O our deontic box-style operator, as in Section 1) – these choices will work for the model in play in the examples.[42] We will actually work with the disjuncts reversed (i.e., using $F \vee E$), to avoid any risk that a reader might think of O as the main connective in $Oq \vee p$:

Examples 2.2. (*i*) Suppose $\mathfrak{M}$ is $\langle W, X, V\rangle$ where $W = \{w_0, w_1, w_2, w_3\}$ with $X = \{w_2, w_3\}$, and $V(p) = \{w_0\}$ while $V(q) = \{w_0, w_2, w_3\}$. Then (relative to $\mathfrak{M}$) $p \vee Oq$ is non-ethical at w_1 because however we vary X to X', calling the model resulting from such a change $\mathcal{M}'$, we have $\mathfrak{M} \models_{w_0} p \vee Oq$ iff $\mathfrak{M}' \models_{w_0} p \vee Oq$, because, since $w_0 \in V(p)$, we shall always have *both* $\mathfrak{M} \models_{w_0} p \vee Oq$ *and* $\mathfrak{M} \models_{w_0} p \vee Oq$, in virtue of the first disjunct's truth at w_0. The same verdicts would be returned for the same reason had the second disjunct been any one of $\neg Oq$, $O\neg q$ or $\neg O\neg q$.

(*ii*) Changing the example to $\neg p \wedge Oq$, we get another formula non-ethical at w_0 because $\mathfrak{M} \models_{w_0} \neg p \wedge Oq$ iff $\mathfrak{M}' \models_{w_0} \neg p \wedge Oq$, however we adjust the set of ideal worlds to obtain $\mathfrak{M}'$, though now this in turn holds because we have *neither* $\mathfrak{M} \models_{w_0} \neg p \wedge Oq$ *nor* $\mathfrak{M}' \models_{w_0} \neg p \wedge Oq$.

(*iii*) Returning to the disjunctive formula in (*i*), but now shifting our attention to w_1, we find that, since $w_1 \notin V(p)$, whether or not $\mathfrak{M} \models_{w_1} p \vee Oq$ depends on whether or not $V(q)$ is a subset of the set of ideal worlds, so since $V(q) \subseteq X$, we do have $\mathfrak{M} \models_{w_1} p \vee Oq$, thanks to the second disjunct, whereas shifting X to $X' = \{w_1, w_2\}$ gives $V(q) \nsubseteq X'$

[42] One may initially think that something like Oq – admittedly not for OA in general (consider $A = q \to q$), but for $A = q$, surely? – should count as ethical at all worlds *in all models*. But no: in models $\langle W, X, V\rangle$ with $V(q) = W$, Oq is true at each $w \in W$ regardless of which subset of X is, so this formula counts as non-ethical. (The corresponding point is made in lines 8–4 from the base of p. 254 in Restall and Russell [2010], whose approach will be related to Karmo's in the following section.) See also Example 3.9. The same goes for the case of $V(q) = \varnothing$, at least if we are restricting attention, as [Restall and Russell, 2010] suggests, models (on frames) for KD45.

and so $\mathfrak{M} \not\models_{w_1} p \vee Oq$. Thus the truth of our disjunction is sensitive to what the set of ideal worlds and the disjunction is accordingly ethical at w_1 in $\mathfrak{M}$. $\lhd$

Ethicality on Karmo's account, as well as being literally contingent or world-relative,[43] is also a property analogous to contingency itself: for contingency proper we have variation depending on which world is under consideration, while for ethicality we must have variation depending on the ethical standard in play. 'Variation' here means in each case that there is *some* way of varying the parameter concerned – world of evaluation or ethical standard – which results in a change in truth-value, not, of course (since there are only two truth-values to go round) that *every* way of varying the given parameter results in such a change (exactly as with contingency itself, indeed).

Karmo then proves (a slightly less formal version of) Proposition 2.4 below, for which we need to introduce the notation $\models_{\mathcal{M}}$ for what is sometimes called the local consequence relation determined by the class $\mathcal{M}$ of models.[44]

Definition 2.3. *$A_1, \ldots, A_n \models_{\mathcal{M}} B$ if and only if for all $\mathfrak{M} \in \mathcal{M}$, where $\mathfrak{M} = \langle W, X, V \rangle$, for all $w \in W$, if $\mathfrak{M} \models_w A_1$, and ..., $\mathfrak{M} \models_w A_n$, then $\mathfrak{M} \models_w B$.*

Karmo's soundness-based version of Hume's Law is then as follows:

Proposition 2.4. *For any formulas $A_1, \ldots, A_n \models_{\mathcal{M}} B$ then for any model $\mathfrak{M} \in \mathcal{M}$ with $\mathfrak{M} = \langle W, X, V \rangle$ and $w \in W$, if $\mathfrak{M} \models_w A_i$ ($i = 1, \ldots, n$) and B is ethical at w in $\mathcal{M}$, then some A_i is ethical at w in $\mathfrak{M}$.*

As with Definition 2.1, of course, Karmo's own formulation makes no reference to models.[45] However, the simple proof Karmo gives of the result carries over to the present formulation without difficulty. Of

[43] No distinction is here intended between these two descriptions, though for other purposes one might want to contrast world-relativity (in the sense of not being world-invariant) with contingency, distinguishing, *à la* McTaggart, a 'B-theory' of modality from an 'A-theory'.

[44] With the notation "$\mathcal{M}$" for a class of models $\mathfrak{M}$, we continue to follow Restall and Russell [2010].

[45] What Karmo has [1988, p. 256], reads as follows: "In general, if sentences S1,..., Sn (where $n > 0$) entail sentence S$(n+1)$, then for any possible world w in which S$(n+1)$ is ethical, if all of S1,..., Sn are true in w, then at least one of S1,..., Sn is ethical in w. (I have added some italics here but resisted the temptation to put the indices into subscript position.) Proposition 2.4 does not include the $n > 0$ condition because it is not needed: we can't have $\models_{\mathcal{M}} B$ (i.e., $\varnothing \models_{\mathcal{M}} B$) for B ethical at a world

course, the approach has not found universal favour and Maitzen [1998] in particular develops several criticisms, to which (as well as the other sources listed in note 1) the interested reader is referred. though here we are more concerned to call attention to connections between the ingredients of Karmo's account and ideas in play elsewhere. We turn in a moment to something of a rediscovery, in (Daniel) Singer [2015], of some of those ingredients – though the recipe in which he combines these ingredient for rescuing a version of Hume's Law turns out not to be quite Karmo's, after illustrating how Karmo's approach handles an objection by Geach, whose own discussion comes closed to anticipating and rejecting that approach – or Shorter-style approaches in general.

Example 2.5. The present example comes from the hard to get hold of Geach [Geach, 1976]. The journal 'Open Mind' was associated with the UK's Open University philosophy course and is not to be confused with the 2017-founded MIT-based cognitive science journal of the same name. Details of the example were included on the second page of Borowski [1976]. Geach is concerned with a version of Hume's Law according to which what he calls morally significant conclusions never follows logically from premises none of which is morally significant, and remarks of his refutation of this principle that "the style of argument is not at all new; I am only refurbishing a weapon already used by Prior, Mavrodes, and others." For Geach's version, we let Y be the last year in which sodomy was illegal in England and are then to consider:

1 Sodomy is either wrong or at least is illegal in England in the year Y.

2 Sodomy is either wrong or at least is illegal in England in the year $Y + 1$.

In reproducing these 'mixed disjunctions', Borowski puts '1967' and '1968' in place of 'Y' and '$Y + 1$', which makes the example easier to think about in the absence of what at least *look like* variables.[46] (This is of some incidental interest because Geach, before introducing 'Y' has

in and model, since B can't be false at any world in any model, so its truth-value is never sensitive to a particular ethical standard (or choice of which worlds are ideal, in the current incarnation of that notion).

[46]They also make the example sound more like something someone might actually say, and it was perhaps with a view to increasing naturalness on this front that Geach included the words "at least" – though this addition adds a complication. Disjunctions in which the second disjunct is prefaced by "at least" or "anyway" often present it as a fallback position introduced in the face of diminishing confidence in asserting the first disjunct outright. Jackson [1987, p. 27], gives the example: "George lives in

said "[t]he English law against sodomy might well change," as though any such change was in the future, as of 1976 – by which time the Sexual Offences Act had already passed into law nine years previously.) Noting that if both 1 and 2 are morally significant then since 1 follows from its second disjunct we have a counterexample to Hume's Law (not that Geach uses any such crass phrase), and that if neither is morally significant, then from 2 together with the negation of its second disjunct we get a counterexample. Accordingly, Geach continues [1976, p. 12]:

> The only hope of saving the 'No *ought* from an *is*' principle is to say that of the pair 1, 2, one is morally significant and the other is not; in fact, that 1 is not morally significant, since it would be inferable from a true premiss that is not morally significant, whereas 2 is morally significant, since from 2 together with a true but not morally significant premiss a morally significant conclusion would follow. This would already be very odd; 1 and 2 differ as regards the date mentioned, and how can that make one morally significant and the other not? But the case again the rule is indeed now much weightier than this. To defend the rule it was necessary to supposed that whether moral significance does or does not attach to a thesis depends not just on the logical structure and sense and force the thesis, but on such grossly empirical matters as the laws recently passed by Parliament. Clearly such considerations cannot affect the application of a proper logical rule.

The defence Geach here envisions (and rejects) on behalf of the differential classification of 1 and 2 is more is in tune with the 'enthymematic' account summarised at the end of Postscript (ii) to Section 3 than with that of Karmo [1988], but let us look at how 1 and 2 fare on the latter's taxonomy. Whether the correct ethical standard endorses the first disjunct of 1 does not affect its truth-value since it is true (in the actual world) in virtue of the truth of its second disjunct however we imagine varying that ethical standard. On the other hand, since the second disjunct of 2 is false (in the actual world), the truth value of 2 depends on

Boston or anyway somewhere in New England," – which would no doubt benefit from some additional punctuation (a comma before "or" at the very least) – and points out that learning that the first disjunct was false would not (by contrast with the case of the second disjunct, equally well introduced by *at least* in place of *anyway*) lead the speaker to retract the assertion. The at least pragmatic failure of commutativity here shows that these are no ordinary disjunctions, and so, not the clear counterexamples they might have seemed to be to 'Hurford's Constraint' (note 56 below). In Geach's case, though, neither disjunct entails (or even 'contextually implies', in the style of Ciardelli and Roelofsen [2017]) the other, so the the order the 'at least' invokes is not one of logical strength; perhaps we are invited to think of the relative seriousness of moral and legal obligations.

the ethical standard. So 1 is non-ethical and 2 is ethical, on Karmo's account. The rhetorical devices Geach employs to make this look like an untenable position are as follows. He introduces the phrase 'morally significant' in such a way that we are not quite clear as whether it is to apply to the basic ethical statements which are indeed settled in world-invariant way by the ethical standard, or to various mixed cases to the ethical/nonethical distinction has to extend to make the treatment dichotomous. In the latter case there seems nothing untoward about a statement's being *de facto* morally significant. Then there is the talk of grossly empirical matters not being the kind of thing that can affect a "proper logical rule," a phrase designed to call to mind rules of inference, perhaps, though Hume's Law is no such thing. Still, Hume's Law does concern itself with the validity of inferences, so perhaps this is not too unfair. We need to recall that Karmo is not emending rather than defending Hume's Law so understood, in replacing the reference to validity with one to soundness – a move Geach's imaginary interlocutor does not quite get round to making – something whose evident dependence on the grossly empirical vicissitudes of life exactly matches that of the contingent taxonomy. ◁

We turn now to Singer [2015] as well as some criticism that has been made of that paper. As already mentioned, Singer (unknowingly) follows Karmo not only in using some contingent ethical/nonethical taxonomy – the basic Shorter strategy – but in drawing this binary distinction in essentially the same way. What he does not do, as we shall see in detail presently, is make the shift from validity to soundness – though unlike Geach's imaginary interlocutor in Example 2.5, he does make a compensatory adjustment to the conclusions of the arguments on which Hume's Law gears. Nor is the vocabulary in which Singer's discussion is couched quite the same as Karmo's, as we have *normative* and *non-normative* rather than *ethical* and *non-ethical*, which is, as mentioned in note 16 above somewhat different, though not in ways that will prevent us from seeing the connection with Karmo's treatment. It is in these rather different terms that Singer [2015, p. 200] presents his formulation of Hume's Law:

> IS-OUGHT GAP: There are no valid arguments from non-normative premises to a relevantly normative conclusion,

and concerning which, where, Singer explains, "a conclusion of an argument is *relevantly normative* when it has substantive normative implications for the possibilities described by the premises (assuming there are

some such possibilities)."[47] We need this to follow the positive proposal, articulated on p. 201:

> Hume gave us an intuitive motivation for IS-OUGHT GAP. Here I take the case one step further by showing that IS-OUGHT GAP, when properly formalized, should be seen as a theorem of normative semantics. If that is correct, the is-ought gap is not subject to Prior's or any other counterexamples. To show this, I assume that normative sentences/utterances are interpreted with respect to points of evaluation that consist of (perhaps among other things) an ordinary possible world and a normative standard.

This will be have a familiar sound to it. It is of course exactly the apparatus we have seen Karmo [1988] introduce to formulate and justify a satisfactory version of Hume's Law (Proposition 2.4 in our somewhat formalized version). Singer remarks that the role of the normative standard – or ethical standard, as Karmo says – can be played by *plans* in the normative semantics (as Singer calls it) in Gibbard [2003]. The happy consilience between Karmo's approach and Gibbard's had been pointed out by James Dreier twenty years before (see [Humberstone, 1996, p. 153]), at which time [Gibbard, 2003] had not appeared but [Gibbard, 1990] had, in which already we see this normative parameter in play, though without the somewhat de-ethicizing expository shift to talk of plans (and without Hume's Law specifically in mind).

How does Singer apply this concept in a repaired version of Hume's Law? First, on p. 202 he introduces the term *norm-invariant* in what a footnote says is to be as understood as in an unpublished paper by Mark Schroeder; this turns out to be simply Karmo's non-ethicality in all worlds. Immediately passing to the 'all worlds' case does not seem promising but instead of what in the following section we shall call de-universalizing the notion w.r.t. worlds (though still quantifying over norms or standards). We have this on p. 203:

[47] In further elaboration of this talk of substantive normative implications from later in [Singer, 2015] (p. 205 to be precise), we have the following, to which I have added italics at one point as a reminder of the – admittedly in need of further precisification – 'guidance' criterion we saw in note 28 had been suggested in Guevara [2008]: "The key claim of IS-OUGHT GAP is this: for arguments from nonnormative premises to a normative conclusion, none of the genuinely normative aspects of the conclusion can be relevant to the possibilities described by the premises. But, since a deductive argument could only help us learn something about how things ought to be insomuch as we accept the premises, any *potential normative guidance* that could be derived from non-normative premises must only apply in possibilities where the premises fail.

> The solution then is to restrict the domain of the is-ought gap to arguments in which the normative aspect of the conclusion is relevant to the possibilities being reasoned about. We can formalize this intuition in our semantic framework easily. To decide whether the conclusion of an argument makes a claim about how things ought to be in the worlds described by the premises, we decide whether the conclusion is norm-invariant when restricted only to the worlds compatible with the premises.

What is it for a world to be 'compatible' with the premises of an argument? This can only mean that we are restricting attention to worlds in which the premises are *true*. So it looks as though we are in for a Shorter-style shift of attention from valid arguments to arguments which are sound in a given world, and are headed towards exactly Karmo's position. But that is not quite how Singer proceeds (still p. 203):

> In our semantics, when the premises are norm-invariant, deciding this is equivalent to deciding whether the conjunction of the conclusion and the premises is norm-invariant. This then is a reformulated version of IS-OUGHT GAP in Gibbard's semantics:
>
> > WORLD-NORM GAP: If $\{P_i\} \vdash C$, each of $\{P_i\}$ is norm-invariant, and $P_1 \wedge P_2 \wedge \ldots$ is satisfiable, then $P_1 \wedge P_2 \wedge \ldots \wedge C$ is norm-invariant.
>
> Intuitively, WORLD-NORM GAP tells us that if the premises of an argument are norm-invariant, then the set of all world-norm pairs compatible with the conclusion and the premises is also norm-invariant. By checking the conclusion conjoined with the premises for norm-invariance, we restrict our attention to only those worlds where the premises are true.

The condition that the conjunction of the premises should be satisfiable has been included, Singer tells us, "to avoid the special case where non-norm-invariant claims follow trivially from contradictory premises." But there is no point in doing this because WORLD-NORM GAP as written is equivalent to the version without the satisfiability condition, as the consequent ("$P_1 \wedge P_2 \wedge \ldots \wedge C$ is norm-invariant.") would automatically be correct whatever the norm-invariance status of what we thought the conclusion of the original argument (namely C) might have been. Eliminating this extra condition brings us closer to a Karmo style formulation (Proposition 2.4), but there is still this awkward feature that we were interested in the status of the argument with premises P_i and conclusion C, and are being told to attend instead to this new argument whose conclusion conjoins the premises with C, a conjunction which is

not in general equivalent to C itself.[48] The interested reader is invited to ponder the persuasiveness of Singer's own explanation as to why this aspect of his approach is, as he goes on (p. 204) to argue, "a feature, not a bug," and to decide whether or not the suggestion is in the end best seen as a rather complicated variation on Karmo's approach.

It is evident that Singer is not himself familiar with Karmo [1988], or he would not have written on the second page of [Singer, 2015] that this was a "rough first pass as the is–ought gap":

> "No normative truth is determined by any non-normative truths,"

and added in a footnote "I formulate the simple version of the claim here in terms of normative and nonnormative truths. It is thus formulated to mirror Hume's talk of propositions, which I take to be bearers of truth-values, though the use of 'truth' in the claim is unnecessary." The 'rough first pass' is indeed rough, but this is a matter of trading in the vague talk of one thing determining another for talk of one statement having another as a consequence, and the unwanted focus, when thus revised, on single-premise arguments; what is not *rough* but – as anyone impressed by Karmo's version of Hume's Law will think – *highly sophisticated* is that the premises in question should be restricted to *truths* for purposes of invoking that law.

In view of the similarities, the contrasting treatment provided by Woods and Maguire [2017] of Karmo and Singer is little surprising. Appended to a sentence (p. 420) which reads "A generation of theorists attempted to characterize the intuitive thesis with increasingly sophisticated logical versions of Hume's dictum," is a footnote describing Pigden [2010] as a "*locus classicus* for these discussions" emphasizing in particular Pigden and Schurz's contributions and adding references to Karmo [1988], Brown [2014], Maguire [2015] as offering "further critique". That is the only reference to Karmo in [Woods and Maguire, 2017], though Singer [2015] comes in for extensive criticism. Before indicating the general drift of that criticism, it should be mentioned that Woods and

[48] A similarly disconcerting shift from one argument to another arises in Borowski [1976; 1980] in which the $\vee$-introduction inference from A to $A \vee B$ figuring in Prior's discussion is replaced by what Borowski calls its (without clearly defining it) that inference's 'canonical form', which is said to be the inference from $A, \neg A$ to B. (The would-be definition of the canonical form of an inference at p. 463 involves talk of replacing its conclusion by "the simplest equivalent proposition whose major connective is implication," as though this conveyed a definite instruction, even once the reference to propositions is replaced by a reference to the kind of thing that might have a major connective, and a specification of what the available logical primitives were taken to be.)

Maguire frequently quote Cuneo and Shafer-Landau [2014] with approval on the idea that there are – or at least we should take seriously the possibility that there are – 'moral fixed points': some substantive moral principles have the status of conceptual truths. Now as Fine on the third page of his [2018] notes, "it is important, if the gap principle is to have any chance of being true, that there be no normatively substantive necessities," since this would make any argument with one of them as a conclusion – certainly if the necessity is conceptual – at least informally valid, even with as uncontroversially non-ethical premises as you like.[49] Of course some of those who would posit such necessities have no intention of saving Hume's Law (the 'gap principle'). Judith Jarvis Thomson suggested that 'Other things being equal, one ought not [to] cause others pain' is a necessary truth and if the necessity is supposed to

[49]This consideration is complicated in Singer's case by the fact that he transforms the original potentially normative conclusion into the conjunction of it with the premises. Fine's reference to 'the gap' here is to the Is–Ought gap rather than to the specific WORLD-NORM GAP of Singer's discussion. Further complicating the discussion are some remarks made by Cuneo and Shafer-Landau, not echoed by Woods and Maguire, that suggest that something's being conceptually necessary does not in fact guarantee that it is true. These conceptual truths, we are told [Cueno and Shafer-Landau, 2014, p. 410*f.*], "hold in virtue of the essences of their constituent concepts" but on p. 413 we read "If the moral fixed points are true, then they are true of conceptual necessity," where the "if" is hard to fathom. There is no Modus Ponens in the offing, so the question of their truth seems to be left open. They follow this conditional formulation with the words: "That is, if we hold certain descriptive information fixed-such as our present human constitution and environment-the concept 'being wrong' is such that it belongs to its essence that, necessarily, if anything falls under the concept 'recreational slaughter' (of a fellow person), then it also falls under it." It is not clear how that something can belong to the essence of a concept conditionally on the state of people and their environment in the way envisaged. The unexpectedly conditional formulation recalls Maitzen's premise in a putative counterexample to Hume's Law in [Maitzen, 1998, p. 354]: "If any ethical sentence is true, torturing babies just for fun is morally wrong." In fact Maitzen gives a disjunctive formulation: Either no ethical sentence, standardly construed, is true, or torturing babies just for fun is morally wrong. (The naturalness of the conditional reformulation here contrasts markedly with that of the Maitzen disjunction featuring in note 57 below, and the text to which that note is appended.) and The other premise is "Some ethical sentences, standardly construed, are true," and the conclusion is the second disjunct of the disjunctive premise. A variation is used in Maitzen [Maitzen, 2010] in which the 'other' premise is instead "At least one (non-negative, atomic) moral proposition is true." Of course, if propositions are to be the common content of logically equivalent sentences, they do not come dressed as negative or non-negative – or even as atomic vs. non-atomic. The difficulties posed by familiar deontic interdefinabilities for any such quasi-syntactic characterization of what it that moral nihilist is *not* to be believe was recognised are recognised in Pigden [2007, p. 452] (though Pigden think they are surmountable and himself seems to want to speak similarly of "non-negative atomic moral propositions").

be conceptual (or analytic), will yield Hume-violating arguments from non-moral premises about actions causing pain to moral conclusions as to their wrongness (other things being equal). References, details and discussion can be found at p. 93 of Sinnott-Armstrong [2000] (or p. 138*ff.* of the book version: [Sinnott-Armstrong, 2006]). Indeed, we already had a foretaste of this from Rynin [1957] in Section 1, note 9. See §2.2 of Sobel [2003] for a discussion of G. E. Moore's view of the (non-analytic) necessity of certain fundamental moral principles.

Woods and Maguire are also sympathetic to Searle [1969]'s famous (putative) derivation – or at least think that it should not be dismissed out of hand – of what someone ought to pay someone five dollars on the basis of premises about what the person said to have the obligation has said, and how linguistic conventions come to constitute this as a promise – premises one would normally take to be on the non-ethical side of the divide, with a conclusion on the ethical side. They accept the formal proof that norm-invariance is passed from the premises to the conclusion of a valid argument (at least in the way Singer makes his case, beefing up the conclusion by conjoining it with the premises, though they could equally well have discussed Karmo's version in which the argument is supposed to be sound and can leave its original conclusion unmolested), and take this to show that norm-invariance is a bad guide to non-ethicality:[50] the premises should be acknowledged to be non-ethical, say Woods and Maguire, even if they are not norm-invariant.[51] It is not entirely clear how neutral on such meta-ethical questions Woods and Maguire are entitled to insist someone drawing an ethical/non-ethical distinction has to be, though. The taxonomy is drawn up with a view to having something like Hume's Law be demonstrably correct with respect to it, so it is not surprising that it won't suit the purposes of those with no interest in salvaging a repaired version of Hume's Law. We should not think of ventures such as Singer's and Karmo's (or Russell and Restall's, reviewed in the following section,

[50]There are some very relaxed formulations in the discussion here: on p. 426 of [Woods and Maguire, 2017] we read "The key fact here is that promissory behavior, given the analytic connection between it and our obligations, is not norm invariant for Searle." (Behaviour itself – as opposed to descriptions or reports of behaviour – is not the kind of thing in the running for being norm invariant.)

[51]Reading [Woods and Maguire, 2017] is further complicated by the fact that Woods and Maguire refer (mid p. 427) to Singer's definition "of ethical facts as *just those which are norm-invariant*" (their italics), which needs "ethical" to be changed to "non-ethical", or to have a "not" inserted before "norm-invariant". This slip occurs several times, including in the subsection titles of 2.3 and 2.4, both of which begin with "Specific Worries about Ethicality as Norm Invariance".

a Postscript to which looks at some aspects of Woods and Maguire's discussion of it) as suasive but rather as explanatory, to use Dummett's terminology (from [Dummett, 1978]) for marking the distinction which in the present instance is that between showing a doubter *that* a version of Hume's Law is correct on the one hand, and showing a potential sympathiser *why* it is correct, on the other.

Woods and Maguire make numerous criticisms of Singer's approach which readily transmute into criticisms of Karmo's, including worrying (p. 430) about allowing w and n to vary independently to arrive at the constellation of all $\langle w, n\rangle$ pairs, which would correspond to having what Karmo calls the ethical standard under consideration be determined in part by the world (and such norms as may be endorsed in it) with which it is paired, in accordance with what our authors call (p. 427) "conventionalist metasemantic views" and feel should be taken seriously. Such a view seems so alien to what Karmo, Gibbard and Singer are doing with the world-norm pairs that it is hard to take to take is seriously in the present setting, though. An ethical standard may dictate that only what is taken to be permissible according to the locally prevailing norms is genuinely permissible, and of course the former will vary – not exactly from world to world since there is not in general one single normative culture (for them to prevail in) per world, but the 'norm' (which is really a normative system or normative standard in Singer's less abbreviated formulations) in a world-norm pair refers to this transcendent set of principles rather than to the any of the prevalent norms by reference to which it fixes the set of actions that are morally permissible.

On their p. 424 Woods and Maguire state the Singer-modified Hume's Law as World-Norm Gap and then parody it with a corresponding World-Octopus Gap.[52] One's immediate reaction to this is perhaps that it is just silly, since the octopus in question is (presumably) already part of the world and so plays no role in the envisaged world-octopus pairs. But it seems that Woods and Maguire's position is that exactly this reaction is not legitimately available, if the account on offer is supposed to be neutral between alternative meta-ethical positions, since it presupposes that the ethical dimension of reality can be segregated

[52]Note the effect achieved by using an animal we find faintly amusing – a bit goofy but potentially endearing – just as with Pigden's references to the hedgehog (note 34). A special gold star should be awarded to Maguire [2015] for managing in a single paper to cite not only Pigden's hedgehog examples but also Dworkin's book *Justice for Hedgehogs* [Dworkin, 2012]. (The book turns out not to be a protest at their ignominious treatment in being so used in the is–ought literature. Instead it discusses justice in general, the title picking up on the hedgehog/fox contrast in intellectual temperament made famous by Isaiah Berlin.)

out from everything, but not every meta-ethical position does allow for such segregation, since there may be conceptual (or even metaphysical) connections between the descriptive and the ethical.[53] If Woods and Maguire had been discussing Karmo himself, they would presumably trace this 'non-neutrality' flaw encapsulated in a single comment, already quoted above, [Karmo, 1988]: "We take it that any possible world can be uniquely picked out with some assignment of truth values to those sentences which the parties to the logical-autonomy-of-ethics debate would unite in calling non-ethical."

Another objection, returning to Singer's discussion, is raised by Woods and Maguire in the following passage, from p. 429 of [Woods and Maguire, 2017]:

> Now, assuming supervenience, let W be the conjunction of all the worldly facts about a possible world w. Let N be some norm that holds in some pair $\langle w, n\rangle$ in our set of factual-ethical pairs. Since we've assumed supervenience, there will be no pair $\langle w, n\rangle$ in our set such that N is not in n. Since this means that N is entailed by W, just as above, either N is descriptive or W is ethical. Neither conclusion is palatable.

First, clarifying the point being made, since we have identified N via the n of an initially given world-norm pair $\langle w, n\rangle$, it would be better to proceed by saying that there will be no pair $\langle w', n\rangle$ in "our set of factual-ethical pairs" (as Woods and Maguire put it), such that N is not in n. So holding n fixed, it would be correct to say that W strictly implies N. But, whatever the Gibbard line on such matters might be, treating this as a potential problem for Karmo's conceptual apparatus (in which n is an ethical standard, oversimplifyingly identified with a set of ideal worlds) N is a necessary truth and so strictly implied by anything you like – but not entailed, since for entailment, the relation we require to hold between the conjunction of the premises of an argument on the one hand, and its conclusion on the other, for that argument to be valid, one needs truth relative to every $\langle w'', n''\rangle$ pair to be preserved by entailment,

[53] Compare the following passage, with which Section 5 of Fine [2018] ends: "A related idea is implicit in the treatment in Gibbard [2003] of worlds as divisible into a descriptive and a purely normative component. But, as we have seen, there is no need for us to go along with this common line of thought. For the truth of the gap principle, as we have formulated it, does not require a clean separation between the normative and descriptive facts; and we may even allow every normative truthmaker will contain a nontrivial descriptive state as a proper part." Other criticisms of Gibbard's proposed semantics for normative language (designed more for grappling with Frege–Geach than for addressing Hume) can be found in pp. 24–26 of Sinnott-Armstrong [2006].

not just for a particular choice of n'' (such as that called by Karmo the *correct* ethical standard).

Fine [2018] takes up with approval several of the points made by Woods and Maguire, his remarks, quoted above, against the 'moral fixed points' endorsed in [Woods and Maguire, 2017], notwithstanding; an example was given in note 53. Fine [2018] makes an elaborate application to the Is–Ought issue, of a truthmaker-based theory of hyperintensionally individuated propositions, about which very little can be said here, where it will serve to round off this section as well as to provide background for a further brief mention in the Postscript. Fine is on the same page as Woods and Maguire in respect of the need not to rule out Searle-style 'promising' arguments, and the account he presents allows one to hold that the conjunction of the premises of such an argument might express a descriptive (or as we would put it, non-ethical) proposition – *Promise*, let's call it, which has as a logical consequence a normative (or 'ethical') conclusion – *ShouldPay*, let's say. The premise is accordingly logically equivalent to the, again normative, $\mathit{Promise} \wedge \mathit{ShouldPay}$, but though equivalent, this is a distinct proposition from *Promise*, enabling us consistently to classify the latter as descriptive even though the former is normative (and these classes are mutually exclusive). Hence the need for a hyperintensional account.[54] The details of the account are somewhat provisional in [Fine, 2018], with Fine frankly noting occasional anomalies and possible repairs.[55]

[54] One might think, for all that has been said about Rynin [1957] here, that when Remnant remarks of Rynin in the opening paragraph of [Remnant, 1959] that "[h]e maintains furthermore that some factual and some moral statements entail each other" what Remnant means is just that according to Rynin, some factual statements entail moral statements and some moral statements entail factual statements, and not literally that according to Rynin there is some pair of statements, one ethical and the other factual, which entail each other. But no, Rynin on p. 317 seems to go somewhat off the rails in the case he makes for exactly this stronger claim. I am not saying that Fine himself is similarly confused with the analogous claim – substituting 'proposition' for 'statement' – but that he is not the first person to have held the moral/normative vs. factual/descriptive distinction to be what we now call hyperintensional.

[55] "There is an awkwardness in the present case which did not arise in the previous case. For in referring to the negative propositions $\neg Q$ and $\neg P$ we have appealed, in effect, to the falsity-makers of P and Q and it would be desirable if we could somehow say what they are in terms of the truth-makers." But at the present stage of the development of Fine's theory, this is not possible. In the middle of p. 564 of [Fine, 2017] Fine writes: "It is important to note that within the present semantics (and this is also true of a number of variants), two formulas A and B may have the same verifiers while $\neg A$ and $\neg B$ do not have the same verifiers. For let A be the formula $p \wedge (q \vee r)$ and B the formula $(p \wedge q) \vee (p \wedge r) \ldots$" (Here I have italicized

Maguire [2015], p. 201, mentions another hyperintensionality example suggested to him by Fine concerning the absorption laws. I will re-notate the statements concerned so as to match the E and F (basic ethical and basic nonethical) notation from Section 1; with that change, what Maguire writes is: "Compare F with $F \vee (F \wedge E)$. They are logically equivalent. F is non-ethical. But in the world in which E obtains, E is one of the grounds of $F \wedge E$ and of $F \vee (F \wedge E)$, which by CONVERSE METAPHYSICAL AUTONOMY is ethical."[56] Note that by contrast with Fine's position in [Fine, 2018], touched on above and in the Postscript below, we seem to have some world-relativity coming in here – not quite of Karmo's kind, though, if E is in the basic ethical category since that would be fixed by the ethical standard and not subject to world-to-world variation. The world-relativity gets into Maguire's version of the truthmaker account because he takes it (p. 196) that

> ... grounding is factive. Non-obtaining facts cannot ground anything. False propositions cannot ground anything.

Although Maguire is officially interested in the metaphysical autonomy rather than the logical autonomy of ethics, in that it is grounding that

the sentence letters which are deliberately left roman in the source, as the italic versions are used as variables over states, sets of which constitute propositions.) In fact, it is not only negation that is a bit peculiar in this semantic account, but even one aspect of conjunction, as is mentioned on the previous page of [Fine, 2017], but which I will put here in terms of the non-linguistic theory of propositions that takes centre stage in [Fine, 2018]: a proposition P and the proposition $P \wedge P$ may be different (albeit equivalent) propositions. The hyperintensionality seems to have got a bit out of control here, though Fine has suggestions as to how to fix this if it should turn out intolerable for some applications of the machinery. The issue with the $P \wedge P$ example arises, as explained in Gautam [1057], from the fact that idempotence for $\wedge$ is not, by contrast with commutativity and associativity, expressed by a linear identity (sometimes called a *regular* linear identity): an equation in which each variable occurs exactly once on each side of the "="; for more information, see note 10 of [Humberstone, 2014].

[56] The content of this principle, given on the preceding page of [Maguire, 2015], is: Any fact partly grounded by an ethical fact is an ethical fact. For the way this is reflected in Fine's account, see the text to which note 61 is appended in the Postscript to this section. One thing making examples involving disjunctions like $F \vee (F \wedge E)$ here hard to think about in general – never mind what F and E are – is that they involve violations of what has come to be called Hurford's Constraint, whether one thinks of this as making for utterance unacceptability (one could not say: sentence ungrammaticality), or just for cognitive processing difficulty; see Ciardelli and Roelofsen [2017] and references there for discussion. (Strictly, in respect of Fine, this issue about the absorption law in question bears not so much on the material in [Fine, 2018], in which $\vee$ and $\wedge$ are operations on propositions, as that in [Fine, 2017], in which they appear as sentence connectives.)

matters rather than entailment, because of this factivity requirement, the account looks close to a Shorter-style account in which entailment is replaced by sound entailment. But I leave the interested reader to survey Maguire's diagnostic discussion (p. 200*f.*) in terms of grounding of Prior's $\vee$-introduction + disjunctive syllogism argument and decide how significantly it differs from the Shorteresque response. The suggestion would be, as with stressing what can be known on the basis of what – see note 33 – what is doing the work is not any deep epistemological (with *knows*) or metaphysical (with *grounds*) issue, but simply the factivity of these notions. We return for a moment to the hyperintensionality issue.

Coincidentally, Maguire quotes a passage from Maitzen [Maitzen, 2010], p. 303 for a different purpose from that for which I would like to draw attention to it – or in fact a slightly longer passage – here; Maitzen is arguing against the world-relativity/contingency aspect of Shorter-inspired positions like Karmo's:

> The contingency thesis makes us implausibly ignorant of the correct classification of disjunctions such as
>
> (GR) Goldbach's Conjecture is true, or Rothenberg's setting his son on fire was morally wrong,
>
> since we don't, and perhaps can't, know the truth-value of one of the disjuncts. The contingency thesis, therefore, implies that we don't, and perhaps can't, know whether GR is moral. Possibly (if implausibly) only its first disjunct is true, in which case GR turns out mathematical and non-moral. Perhaps only its second disjunct is true, in which case GR turns out moral and nonmathematical. Perhaps, instead, the truth of GR is overdetermined by the truth of each disjunct; what is its classification then? It seems odd to say that we can't classify a proposition all of whose components we understand without first knowing which, if any, of those components makes it true.

Whatever the complaint about (GR) is here, it can't be one about the contingency thesis, since the non-moral first disjunct is not contingent. If that disjunct is true, GR is necessarily true and if that disjunct is false (GR) necessarily equivalent to its second disjunct.[57] Since Maitzen

[57]The second disjunct alludes to a real life incident from 1983 in California, in which a Charles Rothenberg deliberately set fire to his six-year-old son David, with near fatal consequences. Maitzen is assuming this will be familiar to readers who accordingly won't be slowed down by the presupposed – or at least backgrounded – non-moral content in this disjunct, and will realise that it is the other disjunct Maitzen has in mind when he says "can't know the truth-value of one of the disjuncts."

seems to want to classify the disjunction differently from its second disjunct whether or not the first disjunct turns out to be false, he is adopting a hyperintensional position of sorts, since even if GR turns out equivalent to its second disjunct, he does not want to be forced to conceded that they agree in respect of ethicality. (I say "of sorts" since the equivalence involved here is necessary equivalence but some may object that this doesn't make is logical equivalence, thereby distinguishing this case from the Fine–Maguire cases of hyperintensionality.)

Section Postscript: Committing Oneself The idea of making a statement which is ethical (in the world in which it is made) seems intimately connected the idea that the class of ethical-in-w statements should be closed under the relation *is entailed by*: this is the converse entailment closure condition in play in the Postscript to Section 1. This is because if assenting to a claim involved committing oneself morally, then assenting to any claim entailing that claim would also involve committing oneself to at least the same extent on at least the same issue as the original claim did. We will look as such special cases of this idea as the case of a conjunct B, entailed by the conjunction $A \wedge B$, so that if B *de facto* (i.e. in the world in question) committed one morally and so counted as ethical in w, then so should the stronger claim with content $A \wedge B$. This was a feature of the world-relative ethical/non-ethical taxonomy in Humberstone [1982a], the details of which (see Postscript (ii) to Section 3) are not important here,[58] but as Karmo explains in these comparative remarks in note 7 of [Karmo, 1988]:

> The present account, unlike Humberstone's, has the appealing feature that if it makes a sentence S ethical at a world w, then it makes the negation of S ethical at w also. On the other hand, Humberstone's account possesses, while the present one lacks, a different appealing feature: if a sentence S entails a sentence S′, and S′ is ethical in w, then so is S. (Consider the conjunction 'Some pigs have wings, and it ought to be the case that all New Zealanders are shot'. On the present account, this sentence will be non-ethical in any world in which no pigs have wings, and this even though

[58] As noted there on p. 475 (and in [Humberstone, 1996]), the account there implausibly classifies every false statement as ethical, though this is not as bad as it might seem given that, in Shorter's wake, only sound arguments are of concern. Dreier notes [2002, p. 247] what may seem a similar if somewhat less serious anomaly for Karmo: "All false statements will have Karmo-moral consequences," though again with this *soundness* perspective in mind one might take a 'who cares?' attitude to the consequences of false statements, and what is ethical in a world, on Karmo's account, is not closed under converse entailment, as the quotation about to be given emphasizes.

> it entails the ethical 'It ought to be the case that all New Zealanders are shot'.)—Exercise: show that a theory having both features will make ethics non-autonomous, in the sense of admitting sound arguments from non-ethical premises to ethical conclusions.

An incidental observation: the 'Exercise' makes it sound as though an account combining the features will have both ethical and non-ethical statements (in world-relative way) but allow conclusions of the former class to be soundly implied by (sets of) premises from the latter class, whereas in fact, as was mentioned in the Postscript to Section 1, and shown in note 27, if either of these classes is non-empty, the other is empty, so no such arguments as seem to be under discussion in the final sentence of the passage just quoted can exist.[59] But let us return to the issue of commitment. The feature of his own account to which Karmo draws adverse attention here, namely that we can have a non-ethical (at w) conjunction even when one of its conjuncts is ethical (at w) is something that can get people – even Karmo himself – a bit confused. Consider, for example, this passage from p. 255 of Karmo's [1988]:

> But suppose Alfie to have issued just one sentence in w, and let this sentence be uncontroversially ethical – let the sentence be 'It ought to be the case that all philosophers are vegetarians'. There will then be two ethical standards E and E′, such that with respect to E, 'Everything that Alfie says is true' is itself true in w, and with respect to E′, 'Everything that Alfie says is true' is itself false in w. (Let E prohibit meat-eating among philosophers, and let E′ refrain from prohibiting meat-eating among philosophers.) No matter what is, in fact, the correct ethical standard—whether E, or E′, or something else altogether—'Everything that Alfie says is true' will be ethical in w. This is an intuitively agreeable result. If Alfie has indeed issued just one sentence, namely 'It ought to be the case that all philosophers are vegetarians', then someone who says that everything that Alfie says is true is himself taking on an ethical commitment (whether he is aware of this or not): the truth value of his comment on Alfie turns on a substantive ethical matter, namely on the permissibility or otherwise of meat-eating among philosophers.

The worry about this passage – from Karmo's own perspective – comes from the talk of commitment. The person who is imagined to claim that everything that Alfie says is true is has supposedly taken on, perhaps unknowingly, an ethical commitment, because the truth-value

[59] Essentially this point was made in [Humberstone, 1996, p. 150, second half].

of the claim about Alfie "turns on a substantive ethical matter, namely on the permissibility or otherwise of meat-eating among philosophers."[60] But wouldn't we regard taking on an ethical commitment as something, whose *truth requires* ethical matters to be a certain way, rather than, less selectively, something whose *truth-value turns on* their being a certain way? With Karmo's footnote 7 example "Some pigs have wings, and it ought to be the case that all New Zealanders are shot", hasn't the envisaged speaker – whether or not pigs fly – taken on an ethical commitment in respect of the treatment of New Zealanders, even if the truth-value of the whole conjunction does not turn on which ethical standard is in play (it being doomed by its first conjunct to falsity in worlds in which pigs do not fly)?

While Karmo's 'Alfie' example is still fresh in our minds, it is only fitting to observe that an essentially similar case has attracted some attention among those unfamiliar with Karmo's discussion:

Example 2.6. Nelson [1995, p. 555] raises, as a serious potential problem for Hume's Law as standardly formulated, the argument from (slightly paraphrasing here) premises (1) Aunt Dahlia believes that Bertie ought to marry Madeline, and (2) All of Aunt Dahlia's beliefs are true, to the conclusion: Bertie ought to marry Madeline. (We meet Aunt Dahlia again in Nelson [2003]. Sinnott-Armstrong [2000] discusses thus current example at length, and seems to think it has something to do with Aunt Dahlia being a reliable authority. But the relevant point is made by changing the premises to "Aunt Dahlia expressed one of her beliefs at time t and what she said then was true" and "Aunt Dahlia expressed the belief at t that Bertie ought to marry Madeline." Or again, change "true" in the new first premise to "false" and change the second premise to "Aunt Dahlia expressed the belief at t that it was not the case that Bertie ought to marry Madeline." Issues of reliability and arguments to authority are beside the point.) Nelson suggests that the conclusion but neither of the premises is ethical, since on the traditional dichotomous and once-and-for-all ethical/non-ethical approach, classifying (2) as ethical would seem bizarre. (For instance, one might add, suppose the premises had been (1′) All of Aunt Dahlia's beliefs are consequences of the proposition that there are rabbits in Australia, (2′) There are rabbits in Australia, and the conclusion had been (2): if (2)

[60] It was because of examples of this kind in Karmo's discussion, that note 16 urged that to avoid entanglement in the issue of whether all ascriptions of truth are somehow normative, that we concentrate on the ethical or moral rather than the normative in any sense broad enough to subsume such ascriptions.

were classified as ethical, we would now have a different counterexample to Hume.) What Nelson does not think to do is what Karmo does, and make the classification of (2) as ethical or otherwise depend on what Aunt Dahlia believes, about which (1), taken as true, gives us crucial information. (Oddly, Wolf [2015, p. 117], cites Pigden, in the introduction to [Pigden, 2010] for such examples, where Pigden explicitly credits them to Nelson [1995], and they were discussed already – the Alfie example – in Karmo [1988], which Wolf discusses elsewhere in [Wolf, 2015].) ◁

Though without the explicit connection to commitment, the idea that an ethical conjunct should make a conjunction ethical surfaces in as different an account as that in Fine [2018], mentioned at the end of the main body of this section, in which the taxonomy applies not to linguistic expressions directly but to propositions conceived as sets of things called *states*, the states which are members of a given proposition being thought of as candidate truth-makers (more specifically, what Fine calls *exact* truth-makers) for that proposition. The states themselves come in two flavours, descriptive and normative, as well standing in a quasi-mereological containment relation to other states. This is from Section 3 of the paper:

> No descriptive state can contain a normative state. It must, in other words, be *purely descriptive*. However, there is no corresponding requirement be purely normative, i.e., contain no non-null descriptive state.[61]

Returning to the linguistic setting and to the commitment issue,

[61] This is Fine's version of the *asymmetry* – as it is called in mid-p. 142 of Humberstone [1996] – needed for commitment-oriented approaches to Hume's Law. But a few lines below the passage from Fine quoted above, the theme of normativity as dominant and descriptivity as recessive arises again at the level of propositions themselves and Fine writes: "We will take a proposition, considered as a set of states, to be *descriptive* if all its member states are descriptive and to be *normative* if at least one of its member states is normative." This is a very different matter, since while at the level of states, contained states, speaking very loosely, behave rather in the manner of conjuncts, all of them required for the state to obtain, whereas at the level of propositions member states behave in the manner of disjuncts, any one of them sufficing for the truth of the proposition concerned. This has nothing to do with whether or not assent registers a normative commitment, and renders the propositional analogues of 'mixed disjunctive' sentences all normative (or 'ethical', in our more customary terminology). Nor would it to say that since the latter are the essentially the negations of the conjunctive cases, they must be treated similarly, since on commitment-oriented accounts (such as that of Humberstone [1982a]: see the end of Postscript (ii) to Section 3, where as well as that, we also have the world-invariant 'partly about M_{eth})' proposal) one does not, *pace* Fine, have closure under negation for the non-basic ethical or non-ethical classes.

Maitzen also has the reaction voiced above, and writing on p. 352 of [Maitzen, 1998], after noting precisely this earlier 'commitment,' mentioning the above passage of Karmo's and its lack of fit with the point about conjunction:

> Consider, for instance, a sentence that conjoins an uncontroversially ethical clause and an uncontroversially non-ethical falsehood: 'Capital punishment is morally wrong, and Montreal is south of New York.' I would classify that sentence as ethical: anyone who assents to the sentence takes on two commitments, one of them ethical. In spite of that commitment, though, Karmo's taxonomy has the sentence come out non-ethical, since the sentence is (actually) false regardless of what the correct ethical standard is.

What is strange, in view of this, is that on p. 350 of the same paper, Maitzen remarks apropos of whether or not "Everything that Alfie says is true" (= A1) is an ethical premise (in the argument from it and "Alfie says that it ought to be the case that everyone is sincere" to the conclusion "It ought to be the case that everyone is sincere") that the answer

> . . . depends on the contingent matter of whether Alfie has, in fact, asserted any ethical sentences: if Alfie *has*, then Al is an ethical premise, since anyone who accepts Al is committed, knowingly or not, to the truth of at least one particular ethical sentence. As Karmo himself puts it, if Alfie has in fact asserted some ethical sentence or other, then Al is an ethical sentence because its truth or falsity 'turns on a substantive ethical matter'.

Despite Karmo's 'off message' remark about commitment, Maitzen's report on Karmo here overlooks the point that that comment was ill-advised precisely because of the actual details of Karmo's treatment: from Alfie's having asserted as many ethical sentences as you like, it does not on that account follow that A1 is an ethical premise, since Alfie may also have made a false non-ethical assertion, in which case the truth-value of A1 is settled – it is false – regardless of the ethical standard in play. This point is illustrated by a minimal variation on Karmo's example "Some pigs have wings, and it ought to be the case that all New Zealanders are shot": just imagine that Alfie makes not this assertion but the two assertions "Some pigs have wings," and "It ought to be the case that all New Zealanders are shot". Thus, this aspect of Karmo's position requires considerable care.[62]

[62] Another issue is also raised by the parenthetical comment in "someone who says

These occasional slips by himself and commentators on it notwithstanding, Karmo seems right to say that his treatment can classify a statement as non-ethical (in a world) despite its entailing – as with a conjunction and either of its conjuncts – something ethical at that world, and that this is a prima facie disadvantage of the treatment. And he is right to say that it offers a compensatory advantage: that the negation of each statement ethical in a given world is again ethical in that world. The disadvantage, as a failure of ethicality to connect with what is seen as morally committal, we saw, led Dreier to discard the Karmo taxonomy, and similar considerations are perhaps at work in the objections raised by Wolf [2015] against Karmo's taxonomy. Wolf invites us (p. 118) to consider the examples:

(BILL) Bill was right to tell the truth about Monica.

(BILL*) Bill was right *not* to tell the truth about Monica.

In particular, we are to consider first the normativity/ethicality status of BILL relative to a world in which Bill lies about Monica. Brushing aside in a footnote the suggestion that what we have here is a case in which "Bill told the truth about Monica" is presupposed by BILL – taking presupposition as the semantic relation last encountered in note 25 – Wolf quickly replaces BILL with the explicitly conjunctive *Bill told the truth about Monica and Bill ought to tell the truth about Monica*, though perhaps that should be "ought to have told" rather than "ought to tell" (the implicature from "ought to have φed" to "did not φ" being readily cancellable). He reminds us that this comes out on Karmo's account as descriptive, rather than normative, since no change

that everything that Alfie says is true is himself taking on an ethical commitment (whether he is aware of this or not). If we were interested in tracking ethical the commitments of a subject, wouldn't it be the subjects beliefs about was the case, rather than what was in fact the case, that were relevant. So argues Dreier [2002]. First (p. 246), he illustrates his dissatisfaction with the failure of the converse entailment closure condition on statements ethical-in-w with a disjunct to disjunction entailment rather than a conjunction to conjunct entailment: "Benito is evil or New Zealand is a Communist Republic" emerging as ethical in the actual world even though it is entailed by its non-ethical second disjunct. (This of course is a version of the motivating consideration – the status of $E \vee F$ in our opening discussion – behind Shorter-style revisions of Hume's Law: the disjunction is entailed, yes, but soundly entailed, no.) Adapting a later example (p. 252) of Dreier's, if we were wanting our taxonomy to mirror ethical commitment and we knew that the speaker believed, firmly though falsely, that New Zealand was a communist republic and asserted the Benito disjunction on that basis, we would no doubt be dissatisfied with its classification as ethical.

in the ethical standard can change its truth-value (given the false first conjunct). Wolf regards this as obviously misclassifying BILL. Here we have the observation, conceded in the second paragraph of the first passage quote from Karmo at the start of this Postscript: it would be nice to have the normatives-in-w closed under converse entailment. The 'commitment' aspect of this is especially in evidence when what's being entailed is one conjunct of a conjunction, since that conjunct is explicitly there in the premise. (A similar sentiment can be found in [Brown, 2015], discussed at the end of Postscript (i) to Section 3.) In fact the normative conjunct in Wolf's conjunctive reformulation of BILL is even more heavily present in BILL itself, since even if one does not have to buy into the presupposition as a semantic (= truth-condition affecting) phenomenon to concede that this normative component is foregrounded in BILL and the descriptive component backgrounded.[63]

Wolf continues (p. 118*f.*):

> Parallel reasons show that BILL* is normative anywhere that Bill doesn't tell the truth. But there is no normatively significant difference between the two—each makes a clear moral evaluation. The only difference is that at some worlds the sentences correctly describe Bill's action and in others they don't. Yet it's difficult to see how this would be relevant to assessing normativity. If it isn't relevant, Karmo's approach doesn't accurately model natural language.
>
> Some might argue that correctly describing Bill's action is normatively relevant, by comparing these cases with Prior's disjunction. Because the disjunction would be descriptive when it describes the facts about tea-drinking correctly, and normative when it doesn't, it gets a mixed treatment, like BILL and BILL*. If it's acceptable for Prior's disjunction to vary with correctness, then perhaps it really is relevant to whether a sentence is normative.
>
> Yet even if we accept the mixed treatment of Prior's disjunction—and we needn't—that would show that correct description is normatively relevant only if correctness does some work toward explaining why we accept different verdicts. Otherwise, correctness might have nothing to do with normativity. Other explanations are plausible: the mixed treatment of Prior's disjunction[64] is tolerable because of what asserting it would commit us

[63] Discussion and references concerning the various contrasts alluded to here can be found in Büring [2007]; alternatively, instead of saying 'foregrounded' – a term the present author regards as preferable to 'focused' since the distinctive aspects of focus particles need not be involved – one could follow Potts [2005] and say that the normative component is *at-issue entailed* by BILL.

[64] That is, the disjunction from [Prior, 1960]: "Either tea-drinking is common in

> to at different worlds. At worlds where we know that tea-drinking is common in England, we can assert the disjunction while denying that New Zealanders ought to be shot. But when we consider worlds where we know that tea-drinking is not common, asserting the disjunction commits us to saying that all New Zealanders should in fact be shot. Karmo's relativity approach reflects the fact that at some worlds we would be committed to obviously normative claims, but not so at other worlds.
>
> Notice there is no similar change in our commitments when we assert BILL or BILL*. Whatever the world, saying that Bill was right to tell the truth about Monica means that Bill ought to tell the truth about Monica. That's a reason for thinking at least some normative sentences stay that way across worlds.

Yes, one could, under pressure from these considerations about commitment, treat the mixed conjunctive cases and the mixed disjunctive cases differently, despite the fact that negation toggles us between the two, as was done in Humberstone [1982a]. That was what Karmo was offering an alternative to, which would preserve closure under negation for the statements deemed ethical in a given world, at the cost of sacrificing closure under converse entailment. These are really not two alternative opinions, but two taxonomies concerning which one might sensibly react as Lewis [1988] does when considering precisifications of the observational/non-observational contrast: you can have a notion of observationality which is closed under converse entailment (so that one observational conjunct observationalizes a conjunction) and you can have a notion of observationality which is closed under negation (so that negating an observational statement gives another observational statement): but if you try for a notion of observationality with both features, things will not go well.[65] These are not world-relative notions in Lewis's case, though they have world-relative analogues, as described in [Hum-

Britain or all New Zealanders ought to be shot."

[65] Another, earlier, venture into philosophical taxonomy prompted, like Lewis's, by logical positivism and the verification principle, not mentioned (though it should have been for the sake of comprehensiveness) in Humberstone [1996] is Morgan [1973], especially as its final paragraph alludes to the normative/non-normative dichotomy. Morgan says on p. 217 "For the sake of this discussion I will assume that we are concerned with a language with the syntactical structure of first order predicate calculus, which may include functions, and which includes the usual connectives", and the mention of function symbols suggests without actually entailing that we are considering first-order logic *with identity*, whose presence would vitiate some of the claim made – such as Lemma 1 on p. 220 which says that the disjunction and conjunction of two formulas sharing no predicate letters are both LC if each of the two formulas is LC, where LC ('logically contingent') formulas are those which are satisfiable and have satisfiable negations. But the disjunction of the predicate-disjoint $Fa \to Fb$

berstone, 1996] and briefly touched on in Postscript (ii) to Section 3 below.

Wolf's own conclusion after presenting difficulties for Karmo's and other responses to Prior's argument is summarised thus:

> The general problem comes from attempting to frame a philosophically significant inference barrier around the distinction between normative and descriptive sentences, which is difficult to pin down. Moore's Law steers clear of these problems because itŠs a semantic barrier: no atomic normative terms are synonymous with any atomic descriptive terms, either directly or by substitution. I think MooreŠs Law can both stand in for the Guillotine and improve on it in an important way.

A similar principle – to the effect that moral concepts cannot be analysed or expressed in entirely nonmoral terms – is called the Moore–Price Law in Sobel [2003], where its logical relations to Hume's Law are examined in some detail. Whether or not Sobel's principle coincides with that favoured by Wolf, his name for it is certainly better, as it does not evoke thoughts of the 'Moore's Law' of computing hardware fame – an unnecessary (and perhaps demeaning) distraction – especially since Wolf doesn't even use the contrasting phrase 'Hume's Law' for what he wants this principle to displace (preferring instead Max Black's terminology: Hume's Guillotine). While Sobel's discussion will not be covered in the present survey, it must be mentioned that it opens with a splendid quotation from Richard Price in which what is mostly known today as Moore's Open Question argument is shown to be have been already alive and well in the eighteenth century.

3 The Restall–Russell Approach

In [2010], Restall and Russell are concerned with classes of models of various types, including in particular models (or interpretations, structures,...) for first order languages, Kripke models for intensional languages, and, potentially, models of other kinds also. What is important

and $\exists y \exists y (x \neq y)$ is not LC even though its disjuncts are. (The criticism of §3 of Brown [Brown, 2014] in note 15 above notwithstanding, Brown is there is alert to the sensitivity of Halldén completeness to the presence or absence of identity. Special attention is paid to the Is–Ought implications of Halldén completeness in §5.1 and Appendix A12 of Schurz [1997].) A more recent discussion prompted by the positivist motivated discussions of demarcating the empirical, which similarly notes the connection with Hume's Law considerations can be found in Diller [2003].

about such models is that they make true, satisfy, or verify certain formulas (or sentences, as we will often say here to follow the usage in [Restall and Russell, 2010]) and not others. If $\mathfrak{M}$ is such a model and A is a formula, we write $\mathfrak{M} \models A$ to indicate that A is true in the model $\mathfrak{M}$. One can make sense of this using the kind of Kripke models we have been mentioning in which a non-empty set W (say) is tupled up with a bit of apparatus for interpreting the intensional vocabulary – a binary accessibility relation in the case of standard Kripke models, or a distinguished subset of W in the case of the simplified Kripke models of the preceding section, or (increasing rather than reducing generality) a function assigning sets of neighbourhoods to the points, etc., – and a valuation function V to assign semantic values to the atomic non-logical expressions (in the propositional case, assigning subsets of W to the sentence letters, though, as explained in note 38, [Restall and Russell, 2010] does not follow this practice). While one can speak of truth in a model so conceived, and this would be taken to amount to truth throughout the model, for many purposes, including Restall and Russell's, it is better to take a Kripke-style model to be a pointed model, in which also a particular element of W is singled out, and truth in the model is taken to amount to truth at that distinguished point (relative to the model concerned).[66] Thus the simplified Kripke models of the previous section above, $\langle W, X, V\rangle$ would become instead $\langle W, X, w, V\rangle$ where $w \in W$ (or, if preferred, $\langle W, X, V, w\rangle$), so that what was formerly written as "$\langle W, X, V\rangle \models_w A$" now becomes "$\langle W, X, w, V\rangle \models A$". In the more general case suited to a normal monomodal logic – as in the case of traditional deontic logic – in place of X here we would have a binary relation on W. Notice that although in the preceding section we found the models without distinguished elements to be easier to use for such purposes as Examples 2.2, in fact Karmo's own informal discussion would favour a formal rendering using the pointed models since it places the correct moral standard, which we can think of as the X of the intended model, and the actual world, which we can (now) think of as the distinguished point of the intended model, completely on a par.

Continuing our exposition of Restall and Russell, suppose, next, that we have a collection $\mathcal{M}$ of such pointed models and a relation $\mathcal{R} \subseteq \mathcal{M} \times \mathcal{M}$. This is not quite the notation used in [Restall and Russell, 2010] but we choose a different font for the relation symbol to minimize the danger of confusing the inter-model relations $\mathcal{R}$ with the intra-model

[66] Pointed models in which the model is generated by the distinguished point are often called rooted models, but this further condition is not imposed here.

accessibility relations. In this notation, Definitions 3 and 4 from [Restall and Russell, 2010] become 3.1(i) and (ii) here, in which $\mathcal{M}$ is a collection of models:

Definitions 3.1. *A formula A of the language interpreted by $\mathcal{M}$*
(i) *$\mathcal{R}$-preserved over $\mathcal{M}$ iff:*

$$\forall \mathfrak{M} \in \mathcal{M}(\mathfrak{M} \models A \Rightarrow \forall \mathfrak{M}' \in \mathcal{M}(\mathfrak{M}\mathcal{R}\mathfrak{M}' \Rightarrow \mathfrak{M}' \models A)).$$

(ii) *$\mathcal{R}$-fragile over $\mathcal{M}$ iff:*

$$\forall \mathfrak{M} \in \mathcal{M}(\mathfrak{M} \models A \Rightarrow \exists \mathfrak{M}' \in \mathcal{M}(\mathfrak{M}\mathcal{R}\mathfrak{M}' \;\&\; \mathfrak{M}' \not\models A)).$$

As the very general terminology suggests, Restall and Russell are not concerned specifically with deontic logic and Hume's Law, but with analogous 'barriers to implication' generally ('inferential barriers' in the terminology of [Humberstone, 1982a] and [Fine, 2018]). These they take pairs of sets of sentences from some language satisfying a condition formulated by reference to the consequence relation $\models_{\mathcal{M}}$ of Definition 2.1 though dropping the quantifier over $w \in W$ and its later subscripted appearances (since we are now working with pointed models or indeed of models as the familiar structures or interpretation in first-order model theory in which there is nothing corresponding to such internal evaluation points for formulas anyway). The condition in question for $\langle \Gamma, \Sigma \rangle$ to be a barrier is that no satisfiable subset of Γ has an element of Σ as a $\models_{\mathcal{M}}$-consequence, where 'satisfiable' means simultaneously true in some $\mathfrak{M} \in \mathcal{M}$: we will call this $\mathcal{M}$-satisfiability for greater explicitness.[67] The main observation is proved without using this terminology however, as Theorem 5.[68] What follows is a mildly reformulated version of this result (also dubbed the 'Barrier Construction Theorem' in [Restall and Russell, 2010, p. 248]:

Proposition 3.2. *For any class of models $\mathcal{M}$, if $A_1, \ldots, A_n \models_{\mathcal{M}} B$, and the set $\{A_1, \ldots, A_n\}$ is $\mathcal{M}$-satisfiable, then there is no $\mathcal{R} \subseteq \mathcal{M} \times \mathcal{M}$ for which all the A_i are $\mathcal{R}$-preserved while B is $\mathcal{R}$-fragile.*

Restall and Russell apply this general result to standard first-order structures with $\mathcal{R}$ as the substructure relation, to conclude that no satisfiable set of substructure-preserved sentences have as a first-order consequence a substructure-fragile sentence, which they regard as vindicating

[67] This is Definition 6 on p. 249 of [Restall and Russell, 2010]; the "$B \in \Gamma$" appearing there is a typo for $B \in \Sigma$.

[68] The point of introducing the notion of barrier is to facilitate is to show – the authors' Theorem 7 – that any barrier thesis can be seen as arising from the preservation and fragility conditions in Theorem 5: a suitable $\mathcal{R}$ can always be found.

a claim of (Bertrand) Russell's to the effect that from no (satisfiable) set of particular premises can one validly infer a universal conclusion,[69] as well as an alethic modal analogue of this which they associated with Kant, in which $\mathcal{R}$ is taken as the submodel relation[70] and for which relation the corresponding notions of preservation and fragility are called modal particularity and modal generality (rather than modal universality, for some reason).[71] There is also a tense-logical application, touched

[69]Restall and Russell in fact say, in the first order case, "semantically particular" and "semantically universal," the adverb being omitted here as *all* of the notions in play in the discussion are characterized semantically. (Russell [2011, p. 150], replaces this adverb with "genuinely".) Restall and Russell, pp. 248 and 250, give the following simple example of a sentence that is neither universal nor particular: $Fa \vee \forall x(Gx)$.

[70]For Restall and Russell, one Kripke model $\mathfrak{M}$ is a *submodel* of another, $\mathfrak{M}^+$ – equivalently $\mathfrak{M}^+$ is an *extension* of the $\mathfrak{M}$ – if they have the same distinguished point, and, using the obvious notation, $W \subseteq W^+$, $R \subseteq R^+$ and V is the restrictions to W ($V(p_i) = V^+(p_i) \cap W$ for each sentence letter p_i). The authors do not require that, similarly, $R = R^+ \cap W \times W$ – i.e. do not require that $\mathfrak{M}$ is the submodel of $\mathfrak{M}^+$ generated by W. That would give a different inter-model relation but would not, as far as I can see, make a difference to which sentences were preserved or fragile w.r.t. the relation in question.

[71]The conspicuously missing reference here would be: Routley and Routley [1969]; cf. also the subsequent discussion in Anderson and Belnap [1975], §§ 5.2.1 and 22.1.2 (the latter by J. A. Coffa). Humberstone [1982] experiments tentatively with the idea of adapting to modal ends, not the "fragile upwards" conception of universality favoured by Restall and Russell, but the "preserved downwards" characterization familiar from the Łos–Tarski Preservation Theorem to the effect that the sentences whose truth is preserved on passage from a first-order structure to an arbitrary substructure thereof are precisely those equivalent to formulas which when written in prenex normal form have all their quantifiers universal. An alethic modal analogue of universality of this kind is called globality in [Humberstone, 1982]. Of course, we again have a Restall–Russell barrier result for the Łoś–Tarski notion of universality: no satisfiable set of such sentences can have a substructure-fragile consequence (though [Restall and Russell, 2010] does not isolate these notions). It does not seem unreasonable as a notion of universality for sentences, which applies to cases such as $\forall x(x = x)$ which do not count as universal in the nomenclature of Restall and Russell. Russell [2011, p. 147] herself mentions the 'upward' version of Łos–Tarski, for formulas with only "$\exists$" in prenex normal form since it is formulas with such equivalents that are preserved under extensions that count as 'particular' in the Restall–Russell classification. ([Russell, 2011] even at one point (p. 146) uses the term *global* – but to characterize Restall–Russell universality rather than Łoś-Tarski universality. The main applications of the Barrier Construction Theorem from [Restall and Russell, 2010] are conveniently summarized in §3 of [Russell, 2011], before the main business is under way: finding an appropriate barrier separating indexical conclusions from the non-indexical premises. The eventual solution is a variation on what Pigden [Pigden, 1989], p. 136*f.* calls the conservativeness of logic and regards as trivializing such barrier theses: this is essentially what the "unless" clause does in Russell's Theorem 5: "No consistent set of constant sentences X entails an indexical sentence A unless X also entails all of A's complete indexical generalisations.")

on in note 73 below, and there are two applications to deontic logic, one of them along the same lines as the alethic modal case and another which is of special current relevance to us.[72] In all cases, since, as Restall and Russell point out, there are formulas that are neither $\mathcal{R}$-preserved over $\mathcal{R}$-fragile, what Proposition 3.2 delivers are Hume-like barrier theses for (setting aside the unsatisfiable cases) threefold rather than twofold classifications: we are in the heart – and perhaps close to the technical summit – of trichotomy territory. So it will take us some further work to see how this connects up with Karmo's dichotomous approach, at least world-by-world, in which the target thesis is a closure-under-consequence condition on the set of nonethical-at-w truths.

Restall and Russell denote by $\langle\!\rangle$ the relation, called by them *normative translation*, defined thus: $\mathfrak{M} \langle\!\rangle \mathfrak{M}'$ iff $\mathfrak{M}$ and $\mathfrak{M}'$ differ at most in respect of their accessibility relations. In the simplified presentation of the models with distinguished subset X this amounts to differing at most over what the distinguished subset is (since the implicit accessibility relation is $W \times X$, where W is the universe of the model).

Proposition 3.2 now tells us that no satisfiable set of $\langle\!\rangle$-preserved formulas can have as a consequence a $\langle\!\rangle$-fragile formula. Of course for a precise statement of the applications of Proposition 3.2, this one and those alluded to in the previous paragraph, we need to know about the underlying $\mathcal{M}$ and for the present application Restall and Russell suggest [2010, p. 253] that we should consider (pointed) models whose accessibility relations are transitive, Euclidean, and serial,[73] which makes $\models_{\mathcal{M}}$

[72] Both deontic applications appear in §5.4, headed 'Normativity I'. §5.5 ('Normativity II'), not discussed here, does not pretend to be anything more than suggestive and envisages an extension relation on 'situations' conceived as a partial version of possible worlds, and of the fragility of normative judgments about them as one passes from a situation to one extending it. The issue seems reminiscent of W. D. Ross's parti-resultant/toti-resultant distinction: additional considerations of any kind, and not just the consideration of additional objects, have the potential to change one's moral assessment of a situation.

[73] They add to this list the condition they call secondary reflexivity, which means that any point accessible to anything is accessible to itself, but this is redundant, following immediately from the Euclidean condition (which says, using S for the accessibility relation as they do, for all model elements x, y, z if Sxy and Sxz, then Syz – so taking z as z we get the redundant condition). The associated deontic schema (the last of those listed on p. 253 and encountered above in Example 1.2) would also be correspondingly also redundant, given the earlier listed $\neg OA \rightarrow O\neg OA$, not that the authors claim otherwise. Singer, discussing Restall and Russell [Singer, 2015, p. 207], writes "They also assume that S is transitive, Euclidean, serial, and secondarily reflexive, though not all of these assumptions are necessary for their proof." Well, in view of the redundancy, not all of these assumptions are necessary for any proof of anything, but when it comes specifically to Restall and Russell's proof(s), *none* of

the local consequence relation of the logic KD45. As is well known, this is also the logic determined by a proper subset of that class of models, namely those $\langle W, S, w, V\rangle$ for which there is X with $\varnothing \neq X \subseteq W$ and $S = W \times X$. As is also well known, we get the same logic by reducing the class of models even further – though this is not something to be exploited here – taking the w-generated submodels of such models, in which case we get the further condition satisfied that $W = X \cup \{w\}$, so that we never have more than one non-ideal world in a model.

(Readers not familiar with tense logic might skip this paragraph.) The simplifications just alluded to assumes that the only modal operators – understood in the broadest sense – are the deontic O, P; we may have additional alethic – or suchlike – operators $\Box, \Diamond$ which, when embedded may direct us from a world in X back out to any point in $W \setminus X$, so we can't afford to throw away all but the initial point of evaluation from among $W \setminus X$. Such additions arise in Example 3.9 below. And in any Kripke model for deontic with an accessibility relation, that relation has a converse and the option arises of introducing operators O^{-1} and P^{-1} which quantify universally and existentially quantify over points to which the current point bears the latter relation as O and P do in the case of the former, validating Hume-inimical 'bridging principle' as it is put in Schurz [1997; 1994], and [2010]: $p \rightarrow OP^{-1}p$. (Note that this is just the familiar tense-logical principle $p \rightarrow GPp$, with P now a past tense $\Diamond$-operator whose consequent put in an appearance in note

these assumptions is necessary since Proposition 3.2 is simply being applied to the case of a particular choice of $\mathcal{M}$ and $\mathcal{R}$, and that general result is indifferent to how $\mathcal{M}$ and $\mathcal{R}$ are chosen. Another strange redundancy occurs in the middle of p. 252 of [Restall and Russell, 2010], where the authors are discussing the accessibility relations of their tense-logical models, and ask us to suppose that this ('earlier than') relation is transitive, irreflexive and antisymmetric. It is already odd to see antisymmetry mentioned in connection with an irreflexive relation, since it is usually cited when one wants to get as close to asymmetry as is possible for a reflexive relation. But since any irreflexive transitive relation is asymmetric, and any asymmetric relation is ('vacuously') antisymmetric, the third condition in their list is redundant either way. (This is not to say that the conditions given suffice for the correctness of the claims they make about them. In mid p. 252 p, Pp, Hp and GPp – in Prior's tense-logical notation – are said to be semantically historic, which is not true in the case of GPp. If the valuation functions, V, V' of two models $\mathfrak{M}, \mathfrak{M}'$ on a frame consisting of the real numbers with 0 as distinguished point, the usual $<$ as accessibility relation, but with $V(p)$ as the set of positive reals and $V'(p)$ as $\varnothing$, then we shall have $\mathfrak{M} \models GPp$ because every point t later than 0 has an earlier point – between t and 0 – verifying p, whereas $\mathfrak{M}' \not\models GPp$ since 0 does have points later than it but p is true at no predecessor of any of them. Yet $\mathfrak{M}$ and $\mathfrak{M}'$ stand in the inter-model relation – V, V' agreeing on the distinguished point and all earlier points – preservation of which makes a sentence semantically historic.)

73, though we could equally have cited the other 'Lemmon bridging axiom', $p \to HFp$). This issue is raised in Example 4.4.29 in Humberstone [2016]. Schurz's own study, as reported in the references just cited, was mainly of mixed deontic–alethic modal logic and so again, does not in general permit of the simplified models even when the deontic fragment is given by KD45.

Let us return to our current concern, which consists in displaying the connections between Russell and Restall's approach and Karmo's.[74] So far, we have seen that both are concentrating on the same class of models. To proceed further it will help to have some terminology more vivid than that used in the opening sentence of this paragraph.

Definition 21 of [Restall and Russell, 2010, p. 254] introduces the term *descriptive* to apply to those sentences which are $\langle\!\rangle$-preserved over the class $\mathcal{M}$, which is a promising start. We then expect a similarly evocative label for the $\langle\!\rangle$-fragile cases. But Restall and Russell's Definition 22, which announces itself as 'Normativity (Sufficient Condition)' tells us just that being $\langle\!\rangle$-fragile is a sufficient condition for counting as a normative sentence. Thus, we don't really have a definition at all.[75] One can see the reason for this: the real definition of normativity comes on the following page, in Definition 23 (see also note 78 below), which is styled simply 'Normativity' and gives as necessary and sufficient for a sentence be normative that it be either $\langle\!\rangle$-fragile or $\Subset$-fragile, where $\Subset$ is the submodel relation (as defined in note 70).[76] Restall and Russell

[74] Imposing this as a condition would also block one of Russell and Restall's proofs, namely that of Lemma 26 (whose content is described in note 79 below).

[75] Though we are at least half way to having one, which is more than can be said for the earlier Definition 2 on p. 247, which purports to define satisfaction (or verification) and reads: Definition 1 (Satisfaction): "Given a formal language L, for each formula A in L, the model $\mathfrak{M}$ will either satisfy that formula (written '$\mathfrak{M} \models A$') or it will not satisfy that formula ('$\mathfrak{M} \not\models A$')." This is just an instance of the law of excluded middle in the metalanguage, and not in the running to be a definition of anything. It's as though the authors had been contemplating the usual kind of inductive definition of $\models$ but decided not to get bogged down in the details, without realising that what they left behind then had no content.

[76] In their summary of this discussion, Woods and Maguire [2017, p. 431] say that Restall and Russell "define *descriptive* sentences as those not ethically fragile in either sense," though, as reported above, [Restall and Russell, 2010]'s Definition 21 defines descriptiveness simply as $\langle\!\rangle$-preservation. And, leaving $\Subset$ out of it, this is not equivalent to the absence of $\langle\!\rangle$-preservation (even if, for satisfiable sentences, it implies it). We can bring in $\Subset$-preservation if we want, by appealing to Lemma 26 of [Restall and Russell, 2010] – see note 79 below – which allows us to rewrite "$\langle\!\rangle$-preserved" to the equivalent "$\langle\!\rangle$-preserved and $\Subset$-preserved," but takes us no closer to something equivalent to "not ethically fragile in either sense", i.e., "not $\langle\!\rangle$-fragile and not $\Subset$-fragile".

never actually introduce a more user-friendly term for $\between$-fragility, using the expression "$\between$-fragile" itself in the course of proving (on p. 256) of what they call the 'normativity formulation' of Hume's Law, the latter being Corollary 25 (from the previous page), which reads: If Σ is a satisfiable set of sentences, each of which is descriptive, and A is normative, then $\Sigma \not\models A$.[77] Since the concept of normativity has been given a disjunctive definition using both $\between$-fragility and $\Subset$-fragility,[78] it seems for present purposes cleaner and more instructive to isolate the $\between$-based concepts both without bringing $\Subset$-fragility into the picture,[79] and define them separately, for which purposes we put an asterisk by the word 'normative' to distinguish it from the $\Subset$-entangled Restall–Russell concept of that name.

Definitions 3.3. (*i*) *A is descriptive iff A is $\between$-preserved over (the current) $\mathcal{M}$.*
(*ii*) *A is normative* iff A is $\between$-fragile over $\mathcal{M}$.*

Then we can extract from the materials of [Restall and Russell, 2010] a direct analogue of the other applications of Proposition 3.2:

Corollary 3.4. *If $A_1, \ldots, A_n \models_{\mathcal{M}} B$, and the set $\{A_1, \ldots, A_n\}$ is a satisfiable set of descriptive sentences, then B is not normative*.*

As with the various $\mathcal{R}$-preservation-vs.-fragile contrasts explicitly in play in [Restall and Russell, 2010], there are sentences which fall into neither category and so we cannot treat Coro. 3.4 as telling us that any

[77] Corollary 24 gave what they call the 'Ought'-formulation of Hume's Law, which applies the $\Subset$-based concepts of normative particularity and normative generality (= preservation and $\Subset$-fragility), which does not bear so directly on our theme, since we are taking a negated *Ought*-judgment to be just as much a potential ethical conclusion as an unnegated *Ought*-judgment. See note 19 and the text to which it is appended, above.

[78] Restall and Russell write "$\Supset$-fragile" in Definitions 19 and 20 on p. 254, and in Definition 23 on p. 255 (where also the "$\mathfrak{M}' \models A$" in the third line is a typo for "$\mathfrak{M}' \not\models A$"), which is understandable since it is fragility travelling upward to extensions, but by the letter of the generic definitions of $\mathcal{R}$-preservation and $\mathcal{R}$-fragility in Defs. 3 and 4 on p. 248, reproduced in our Definitions 3.1(*i*) and (*ii*), the correct formulation demands $\Subset$-fragility, and, where they write "$\Supset$-preservation", $\Subset$-preservation. The pre-hyphenated inter-model relation symbols in Restall and Russell's Definitions 8 and 9 (p. 250), 11 and 12 (p. 251) are all the wrong way round for the same reason. Fortunately, since we are concentrating on the symmetric $\between$, no such correction is required in the cases of present interest.

[79] The fact notwithstanding that, according to the interesting Lemma 26 of [Restall and Russell, 2010], all descriptive sentences are normatively particular – i.e., $\between$-preservation implies $\Subset$-preservation.

satisfiable set of descriptive premises has only descriptive conclusions as consequences. For instance, $p \vee Op$ is a consequence of the descriptive p but is not itself descriptive since if we take $\mathfrak{M}$ with $\mathfrak{M} \not\models p$ although $\mathfrak{M} \models Op$, we have $\mathfrak{M}$ verifying the disjunction despite having a $\langle\!\rangle$-related $\mathfrak{M}'$ which consists in adding the distinguished point to the set of ideal worlds of $\mathfrak{M}$, with the effect that $\mathfrak{M}' \not\models p \vee Op$. (Note that the 'translation' relation $\langle\!\rangle$ does not change the distinguished point of these pointed models.)[80] Nor is $p \vee Op$ normative*: $\langle\!\rangle$-fragility is out for the same reasons as in the alethic and quantificational cases mentioned in the preceding note: no adjustments to the set of ideal worlds (or the accessibility relation) will take $\mathfrak{M}$ to a $\mathfrak{M}'$ with $\mathfrak{M}' \not\models p \vee Op$, if the reason we have $\mathfrak{M}' \models p \vee Op$ is that $\mathfrak{M}' \models p \vee Op$. Accordingly, as already stressed, what we get is not a Humean dichotomy, but a quasi-Humean trichotomy.

In view of such considerations, it is somewhat surprising to read Mares [2010, p. 283] saying in what purports to be a summary of the Restall–Russell account "A formula is *fragile* if and only if it is *not* preserved." Even if Restall and Russell had not explicitly disavowed any such claim (as they do: see note 69 above, for instance) – since, as one can see from Definitions 3.1, $\mathcal{R}$-preservation and $\mathcal{R}$-fragility are respectively $\forall\forall$ and $\forall\exists$ notions, it would only be under exceptional circumstances that they could end up being complementary. Probably what Mares was thinking of was not the properties of sentences or formulas of being $\mathcal{R}$-preserved or being $\mathcal{R}$-fragile, but the relations between formulas and models that results from removing the initial universal quantifier "$\forall\mathfrak{M}$" from the Definitions 3.1(i) and (ii) – or more precisely from the *definientia* involved (i.e., the parts after the "iff"); this would turn the definitions into (i) and (ii) here:

Definitions 3.5. *For any class of models $\mathcal{M}$ and any sentence A which can be interpreted in $\mathcal{M}$:*

(i) *A is $\mathcal{R}$-preserved from $\mathfrak{M} \in \mathcal{M}$ (over $\mathcal{M}$) iff*

$$\mathfrak{M} \models A \Rightarrow \forall\mathfrak{M}' \in \mathcal{M}(\mathfrak{M}\mathcal{R}\mathfrak{M}' \Rightarrow \mathfrak{M}' \models A).$$

(ii) *A is $\mathcal{R}$-fragile from $\mathfrak{M} \in \mathcal{M}$ (over $\mathcal{M}$) iff*

[80] We could have used instead the case of $p \vee Oq$ to illustrate this point, with a suitable choice of $V(q)$, but give the present example because of its novelty as compared with the universal and modally general examples from Restall and Russell: in those cases the point could not have been made with $Fa \vee \forall x(Fx)$ or $p \vee \Box p$, because these disjunctions are equivalent to their first disjuncts.

$$\mathfrak{M} \models A \Rightarrow \exists \mathfrak{M}' \in \mathcal{M}(\mathfrak{M} \mathcal{R} \mathfrak{M}' \;\&\; \mathfrak{M}' \not\models A).$$

Since the parts following the "$\mathfrak{M} \models A \Rightarrow$" in the defining conditions in (i) and (ii) here are equivalent to each other's negations, this would then give a two-block partition of the formulas true in $\mathfrak{M}$; such truths, that is, would then fall into exactly one of the categories: $\mathcal{R}$-preserved from $\mathfrak{M}$, $\mathcal{R}$-fragile from $\mathfrak{M}$, and Mares's comment so reinterpreted would be correct. Further, since we are concentrating on the truths (in some pointed model, in the deontic application of this), we might well be in business for some kind of Shorter-inspired *soundness* version of Hume's Law. Before pondering the deontic/ethical case specifically, though, let us state the general (and easily proved) 'Shorterized' version of Restall and Russell; here $\mathcal{M}$ and $A_1, \ldots, A_n, B$ are related as are $\mathcal{M}$ and A in Definitions 3.5:

Proposition 3.6. *If $A_1, \ldots, A_n \models_{\mathcal{M}} B$, and for $\mathfrak{M} \in \mathcal{M}$ we have $\mathfrak{M} \models A_i$ (each $i = 1, \ldots, n$), then there is no $\mathcal{R} \subseteq \mathcal{M} \times \mathcal{M}$ for which all the A_i are $\mathcal{R}$-preserved from $\mathfrak{M}$ while B is $\mathcal{R}$-fragile from $\mathfrak{M}$.*

Now specializing the discussion back to the ethical case and taking $\mathcal{R}$ as Restall and Russell's $\langle\!\rangle$, we note that in terms of the unpointed models $\langle W, X, V\rangle$ in play in Definition 2.1, A's being non-ethical at $w \in W$ in such a model amounts A's having the same truth-value at w in all of the models $\langle W, X', V\rangle$ varying the ethical standard X. Transferring this across to the framework of Restall and Russell, but with the de-universalized model-specific (or model relativized) notions of Definitions 3.5 in place, we get that being non-ethical in the pointed model, $\mathfrak{M} = \langle W, X, w, V\rangle$ amounts to A's having the same truth-value in all $\mathfrak{M}'$ which are $\langle\!\rangle$-related to $\mathfrak{M}$. But this isn't quite what Definition 3.5(i) itself says being $\mathcal{R}$-preserved from $\mathfrak{M}$ consists in, when $\mathcal{R}$ is taken as $\langle\!\rangle$. Rather, being $\langle\!\rangle$-preserved from $\mathfrak{M}$ is a matter of being true in all $\mathfrak{M}'$ which are $\langle\!\rangle$-related to $\mathfrak{M}$ if A is true in $\mathfrak{M}$, and this does not address the question of what happens if $\mathfrak{M} \not\models A$.

To arrive, as we shall after Definition 3.8(ii) below, at a de-universalized Restall–Russell formulation matching Karmo's, we need to back up for a moment with a few general remarks about the general process involved. Consider two first-order sentences:

$$\forall x \forall y(Sxy \to Syx) \qquad\qquad \forall x \forall y(Syx \to Sxy)$$

They are just two ways of saying that (the binary relation interpreting) S is symmetric. Removing from each of them the initial universal

quantifier binding x gives two non-equivalent conditions for an individual (value of) x to satisfy, which we could denote by lambda expressions in an obvious way: $\lambda x \forall y(Sxy \rightarrow Syx)$ – standing for the property of being, as we might say, an S-reciprocatee, and $\lambda x \forall y(Syx \rightarrow Sxy)$ – for the property of being an S-reciprocater.[81] This is illustrates the fact that there is no such thing as *the property predicated of everything* by a closed sentence of the form $\forall x(\varphi(x))$, if we want the property concerned not to depend on the syntactic shape of the sentence but to be shared by all logically equivalent sentences, in much the same way as there is no such thing as *the* property predicated of a by a sentence $\phi(a)$ (a being an individual constant), which was illustrated in [Humberstone, 2000] for the case of $\phi(a) = Fa$ (F a monadic predicate letter).

Intermission. Given that the example just given by 'de-universalizing' the claim that a binary relation was symmetric – though such a description must be understood to denote removing only the outermost universal quantifier, rather than all of them – one may wonder if a similar possibility arises with a universally quantified monadic predication. The answer is that it does:

Example 3.7. Take the sentence $\forall x(Fx)$, which says that everything satisfies the condition $\lambda x(Fx)$. Can we find a condition which is not equivalent to this which is such that the sentence that everything satisfies that other condition is equivalent to $\forall x(Fx)$? In classical first-order logic with identity certainly we can. On which comes to mind is the following:

$$\lambda x(\exists y(Fy) \wedge \forall z(z \neq x \rightarrow Fz)).$$

The reader is invited to that putting $\forall$ in place of λ gives an equivalent of $\forall x(Fx)$, while predicating the two properties involved of a given individual (we again use the constant a) gives the non-equivalent Fa and $\exists y(Fy) \wedge \forall z(z \neq a \rightarrow Fz)$. $\lhd$

It would be interesting to have some idea of the what the inverse image of a given universal formula is, in the sense of knowing what the set of open formulas $\phi(x)$ (as we may as well write in place of "$\lambda x(\varphi(x))$") looks like for a given closed universal formula $\forall x(\phi(x))$, all of them equivalent to that $\forall$-formula. A similar line of enquiry is opened up for the case of $\Box$-formulas in modal logic in [Humberstone, 2013], where of course the set of formulas whose necessitations are equivalent to a

[81] Points in a Kripke frame with S as accessibility relation are called 1-symmetric and 2-symmetric respectively in [Humberstone, 2016], p. 188*ff.* with a similar – though three-way – distinction in the case of transitivity (p. 185*ff.*).

given $\Box$-formula will vary from one to another modal logic. ***End of Intermission.***

Here we concern ourselves with some specific cases of de-universalizing bearing on the Hume's Law theme, another such case being addressed in Postscript (ii) to this section. We continue to think of de-universalizing as a syntactic process of removing the main universal quantifier from an $\forall$-formula (binding with a 'λ', if desired, the variable thus freed[82]): applying this syntactic operation to all formulas equivalent to the that formula yields the members of its inverse $\forall$-image. So for a closer *rapprochement* with Karmo, we need go to back and replace the "preserves $\mathcal{R}$" idea with something that alludes to both preserving and reflecting (as it is sometimes put) the property of being true in a model. We will use the word *copied* for this stronger property. The *definiens* in Definition 3.1(i) for $\mathcal{R}$-preservation can be re-expressed, after shifting a quantifier and 'permuting antecedents' so that it looks like this:

$$\forall \mathfrak{M}, \mathfrak{M}' \in \mathcal{M}(\mathfrak{M}\mathcal{R}\mathfrak{M}' \Rightarrow (\mathfrak{M} \models A \Rightarrow \mathfrak{M}' \models A)).$$

So all we have to do is to boost the last "$\Rightarrow$" to a "$\Leftrightarrow$" to get a Restall–Russell style condition (3.8(i) here) and then de-universalize again (3.8(ii)) to get the model-specific version:

Definitions 3.8. (i) *A is $\mathcal{R}$-copied over $\mathcal{M}$ iff:*

$$\forall \mathfrak{M}, \mathfrak{M}' \in \mathcal{M}(\mathfrak{M}\mathcal{R}\mathfrak{M}' \Rightarrow (\mathfrak{M} \models A \Leftrightarrow \mathfrak{M}' \models A));$$

(ii) *A is $\mathcal{R}$-copied from $\mathfrak{M}$ (over $\mathcal{M}$) iff:*

$$\mathfrak{M}' \in \mathcal{M}(\mathfrak{M}\mathcal{R}\mathfrak{M}' \Rightarrow (\mathfrak{M} \models A \Leftrightarrow \mathfrak{M}' \models A)).$$

In general, being $\mathcal{R}$-copied is a very different property of formulas from being $\mathcal{R}$-preserved, so there may be a feeling that we are relying only on a loose analogy in connecting Restall and Russell's approach to Karmo's, but note that for symmetric $\mathcal{R}$, being $\mathcal{R}$-copied and being $\mathcal{R}$-preserved completely coincide, and $\between$ is a symmetric relation (indeed, an equivalence relation). So if Restall and Russell had chosen simply to address Hume's Law in [Restall and Russell, 2010] and to do so in the pure $\between$-based setting, they could equally well have done so by defining the descriptive sentences to be those $\between$-copied over the relevant $\mathcal{M}$ as they do by defining them to be those sentences which are $\between$-preserved over class $\mathcal{M}$: these are just two characterizations of the same set of

[82] What if no occurrences of the quantified variable *are* thus freed? It is perhaps not immediately clear whether vacuous universal quantifiers should be excluded here.

sentences.[83] As with the cases touched on above, de-universalizing gives non-equivalent results and in particular de-universalizing in the $\wr$-copied case gives Definition 3.8(ii), yielding Karmo-style non-ethicality in $\mathfrak{M}$ (or: at w in the 'unpointed' reduct of the pointed model $\mathfrak{M}$, where w is the distinguished point of $\mathfrak{M}$). Since Karmo's discussion has a very clear conception of an intended model (with the actual world as distinguished point and the correct ethical standard as the ethical standard in place), de-universalizing the general notion to focus on relativity to this intended model, $\mathfrak{M}^*$, say is close to irresistible: ethicality at the actual world is $\wr$-fragility from $\mathfrak{M}^*$. With these move, then, we remove the appearance of a discontinuity between Karmo's treatment and the de-universalized model-relative version of the Restall–Russell account.

Corresponding to what was described after Definition 2.1 as a more direct adaptation of one of Karmo's formulations – though negating it, since it is now *non*-ethicality that is at issue, instead of defining this model-relative notion of descriptiveness or non-ethicality by saying that A has this property relative to $\mathfrak{M}$ just in case:

$$\forall \mathfrak{M}' \in \mathcal{M}(\mathfrak{M} \wr \mathfrak{M}' \Rightarrow (\mathfrak{M} \models A \Leftrightarrow \mathfrak{M}' \models A)),$$

we can equivalently put this as follows, for the reasons given in the discussion after Definition 2.1:

$$\forall \mathfrak{M}', \mathfrak{M}'' \in \mathcal{M}((\mathfrak{M} \wr \mathfrak{M}' \,\&\, \mathfrak{M} \wr \mathfrak{M}'') \Rightarrow (\mathfrak{M}' \models A \Leftrightarrow \mathfrak{M}'' \models A)).$$

Aside from considering such de-universalized versions of the Restall and Russell concepts to make contact with Karmo's approach to Hume's Law, it is worth spending a moment on their role in [Restall and Russell, 2010] without reference to Karmo. In the first place, Restall and Russell in fact help themselves occasionally to these model-relative notions

[83] It is therefore surprising to read Russell [2011] in Remark 1 on p. 157 contrasting her approach there with the earlier Barrier Theorem work: "Instead of looking at whether the *truth* of a sentence is always preserved over changes, the definitions of constant and indexical sentences look at whether *truth-value* is preserved over changes." But this is no contrast at all when the changes are all reversible, as changes to $\mathcal{R}$-related structures for symmetric $\mathcal{R}$ are – the structures here being models paired with contexts and $\mathcal{R}$ relates any two agreeing on the model component of the pair: a symmetric relation. Indeed on p. 159 Russell writes "On our new approach to the indexical barrier theorem, the relation remains symmetric," discussing the bearing of this on another aspect of her treatment: whether the barrier operates in the reverse direction also – not quite the same issue. The treatment in [Russell, 2011] is indeed a new departure, since while the constant sentences are those preserved by the $\mathcal{R}$ just mentioned, the indexicals comprise simply just the complementary class, rather than being given the fragility treatment (and there is consequently another wrinkle in the treatment – mentioned at the end of note 71).

without explicit acknowledgement, for the sake of heuristic remarks. On p. 248, the authors are considering the inter-model relation $\mathcal{R}$ (as we shall write it, though they write simply 'R') as the substructure relation as the substructure relation, writing:

> Take the example of $Fa \vee \forall x(Gx)$. This is sometimes $\mathcal{R}$-preserved (if you have a model in which Fa is satisfied, $Fa \vee \forall x(Gx)$ is satisfied in any extension of it). However, it is sometimes not (take a model in which Fa is false, but $\forall x(Fx)$ is true – extend it to a model in which G is fails of some objects).[84]

Of course, there isn't literally such a thing as being "sometimes $\mathcal{R}$-preserved"; the more careful way of saying this is that Fa is true in a model, then $Fa \vee \forall x(Gx)$ is substructure-preserved from that model, whereas if Fa is false, it is not. (Note the similarity to the deontic 'mixed disjunctions' of Prior's argument.) Similarly, Wolf [2015, p. 119] says at the start of his summary of what he calls the *fragility approach* of Restall and Russell that "it designates a sentence as normative if just in case there is at least one modal where replacements and additions to the set of satisfactory worlds changes its truth-value", so here was have lost the $\forall$ from the authors' official $\forall\exists$ definition and are working with *fragility from* a given model – essentially, in other words, with Karmo's ethicality at a world in a model (not fussing here too much about the "replacements and additions" formulation and taking it to amount to "changes").

Another issue with which the more refined concepts introduced in Definitions 3.5 (or the similarly model-relative variant in definition 3.8 (ii)) promise assistance is in dealing with an objection to [Restall and Russell, 2010] from Vranas [2010], p. 263. Vranas puts his objection in terms of Restall–Russell normativity rather than what was called normativity* in Definition 3.3(ii), but here, to avoid complications, we present it in the latter (purely $\Diamond$-involving) concept:

Example 3.9. Suppose we have an alethic modal operator present $\Box$ interpreted in the deontic models under consideration by Restall and Russell, though (as Vranas acknowledges) not present in the object language they use such models to interpret, and we interpret it by universal

[84] I have changed the notation to match that in use here turn the authors' "R" becoming "$\mathcal{R}$" and their "$(\forall x)Fx$" becoming "$\forall x(Fx)$". Before the passage quoted here, Restall and Russell describe the inter-model relation involved as the relation of model extension, rather than substructure. This is the mistake mentioned in note 78 surfacing again.

quantification over all the model elements (not just the ideal points, as with O). Then, contrary to the application of the Barrier Construction Theorem – Proposition 3.2 in our development – the valid inference from $\Box p$ to Op takes us from the descriptive to the normative*, i.e. from the $\langle\!\rangle$-preserved to the $\langle\!\rangle$-fragile. (Vranas is concerned with the passage from the descriptive to the normative – no asterisk – the latter concept involving also $\Subset$-fragility, and has a diagnosis of what goes wrong involving this aspect of the case, but let's stick with the simple purely $\langle\!\rangle$-based version.) $\Box p$ is certainly $\langle\!\rangle$-preserved: shifting around the set of ideal worlds does nothing to change the universe (W) of the models so if a model verifies $\Box p$ before the shift (or the translation, to use Restall and Russell's favoured geometric metaphor), the same will be so after the shift. But Op is not $\langle\!\rangle$-fragile, as we saw – and observed that Restall and Russell had already seen – in note 42 (where the example was actually Oq). That much can be said in terms of the concepts officially available in [Restall and Russell, 2010], where the issue is raised on p. 254 with the words "Oddly enough, important normatively general sentences such as Op are not $\langle\!\rangle$-fragile," the explanation being as given in note 42 above, which does not entirely deal with the "oddly enough" aspect of the situation. This issue is touched on in Schurz's comments on Restall and Russell (and Vranas) [Schurz, 2010a], p. 271, with the observation that if you want to the implication from $\Box p$ to Op to be respected by your logic, you need to restrict the class of models $\mathcal{M}$ for which you are taking the consequence relation $\models_{\mathcal{M}}$ as your logic, you have insist that the $\Box$-pertinent alternatives include all the O-pertinent alternatives (in the simplified case: that $W \supseteq X$) and you lose $\langle\!\rangle$-fragility, whereas if you want $\langle\!\rangle$-fragility you need to exclude models meeting this condition and then your consequence relation will not deliver Op as a consequence of $\Box p$. One can make a somewhat finer-grained response, though, with the model-relative notions to hand: suppose $\mathfrak{M} \models \Box p$; then we know not just that Op is not $\langle\!\rangle$-fragile – a general claim – but that, though this is not a $\langle\!\rangle$-preserved formula, it is $\langle\!\rangle$-*preserved from* $\mathfrak{M}$. $\lhd$

The implication from $\Box p$ to Op, or more generally from $\Box A$ to OA under discussion in Example 3.9 has been the subject of strong hostility – with objections to the provability of such things as $O(p \vee \neg p)$ in even monomodal deontic logic (i.e. without an additional primitive $\Box$). Pertinent quotations from Jonathan Harrison and Chares Pigden, as well as pointers to suggested remedies, can be found in Remark 4.4.9 in Humberstone [2016]. The implication is often called Must-implies-Ought by analogy with Ought-implies-Can, but this is potentially confusing be-

cause there is also the deontic 'must' to contend with[85] – which is what is meant in the title of Vranas [2018], as well as that of Jones and Pörn [1986] – one of the places just alluded to as offering a remedy, in fact, for the deontic operator (written as 'Ought') defined at the top of their p. 92.

This completes our guided tour through the recent post-Prior literature. In Section 2 we found aspects of Karmo [1988], developing a Shorter-style response to the difficulties Prior raised for Hume's Law by working with soundness and a world-relative dichotomous taxonomy, resurfacing in Singer [2015], though we also sampled criticisms of Karmo and of Singer by Maitzen and by Woods and Maguire, respectively, and briefly touched on Fine [2018]'s distinctive hyperintensional approach to the issues. (Some indication of how Fine approaches Hume's Law itself is given at the end of Postscript (i) to this section.) In this section we have seen that treatment of Hume's Law by Restall and Russell as a special case of their general account of barrier theses in various areas. While, again, the reception has not been uniformly favourable, we have concentrated less on the criticisms than on the connections which arise with Karmo's approach in particular, once their key concepts (of $\langle\!\rangle$-preservation and $\langle\!\rangle$-fragility) are simplified in a certain way – de-universalized, as we put it; further connections with work of David Lewis come up in Postscript (ii) below. Of course, the views of numerous others – and not even just those named in the opening paragraph of Section 1 have also been brought into the mix, but that will do by way of a concluding paragraph.

Section Postscript (i): Woods and Maguire on Restall and Russell We pick up the discussion in Section 3 [Woods and Maguire, 2017] of Restall and Russell from note 76. The second paragraph of [Woods and Maguire, 2017]'s §3.1 includes a proof of what looks vaguely like the main result in [Restall and Russell, 2010], their Barrier Construction Theorem (Theorem 5 in their paper, a formulation of which appeared as Proposition 3.2 here), though on closer inspection turns out not to be.

Recall that according to that result for any class of models $\mathcal{M}$, if B is a semantic consequence of $A_1, \ldots, A_n$ over $\mathcal{M}$, then for no $\mathcal{R} \subseteq \mathcal{M} \times \mathcal{M}$ can it be that all the A_i are $\mathcal{R}$-preserved while B is $\mathcal{R}$-fragile. The definition of $\mathcal{R}$-fragility given by Woods and Maguire in the second paragraph of 3.1 of [Woods and Maguire, 2017] correctly captures the notion in play in Restall and Russell's discussion, but they do introduce

[85] To say nothing of the epistemic 'must': "It must have rained in the night."

the concept of being $\mathcal{R}$-preserved here, instead defining a sentence to be $\mathcal{R}$-*stable* iff it is not $\mathcal{R}$-fragile. This is a clue that we are not going to be shown Restall and Russell's main result, or a simplified version (with the same range of application), in case that is what Woods and Maguire hoped to do, in avoiding the concept of $\mathcal{R}$-preservation. What claim to be proving is the following on p. 431 of Woods and Maguire [2017]:

"An $\mathcal{R}$-stable sentence does not imply any sentence that is $\mathcal{R}$-fragile."

Compare the Restall and Russell version, re-worded into talk of implication: "A satisfiable set of $\mathcal{R}$-preserved sentences does not imply any sentence that is $\mathcal{R}$-fragile."

Concerning their own claim, Woods and Maguire say "The proof is easy." There is indeed a simple proof, but Woods and Maguire's proof is not easy to follow at all; comments indicating why are included here in doubled brackets; the authors' use of φ, ψ as schematic letters is followed to facilitate checking that the source text has been accurately reproduced here (except for the R which here appears – as above – as $\mathcal{R}$):

> Let φ be $\mathcal{R}$-stable and ψ be $\mathcal{R}$-fragile. $\mathcal{R}$-stable sentences are consistent by definition. [[That step is correct, since being stable means that there is a model verifying the sentence – so the sentence is consistent – and every model it bears the relation $\mathcal{R}$ to also verifies the sentence.]] If φ and ψ are not jointly inconsistent [[that should be "not jointly *consistent*"]], then any model of φ witnesses the failure of the implication of ψ from φ. If they're jointly consistent, we have a model $\mathfrak{M}$ of both φ and ψ. Since ψ is $\mathcal{R}$-fragile, we can extend [[here meaning: pass to some $\mathcal{R}$-related model]] the model to some $\mathfrak{M}^*$ where ψ is false. Since $\mathcal{R}$-stable sentences true in $\mathfrak{M}$ are true in $\mathfrak{M}^*$, φ is true in $\mathfrak{M}^*$ and we have our counterexample. [[If φ had been assumed to be $\mathcal{R}$-preserved, we could argue that way – "$\mathcal{R}$-stable sentences true in $\mathfrak{M}$ are true in $\mathfrak{M}^*$" would follow, but not with the mere assumption of $\mathcal{R}$-stability. All the latter means is that there is some model, $\mathfrak{M}_0$, say, verifying φ with every model $\mathcal{R}$-related to $\mathfrak{M}_0$, also verifying φ. But who says that the $\mathfrak{M}$ introduced in the course of the proof to be some model verifying both φ and ψ (assumed consistent) is such an $\mathfrak{M}_0$, all models $\mathcal{R}$-related to which continue to verify ϕ?]]

In short, this would-be proof of a result which isn't Restall and Russell's anyway, is not a great success, though the result in question is not in doubt. To see that, for the record, let us pick up the proof from the correct initial step, inferring from the $\mathcal{R}$ stability of φ – an $\exists\forall$ property, since $\mathcal{R}$-fragility is an $\forall\exists$-property – that there is a model $\mathfrak{M}$ such that

(1) $\mathfrak{M} \models \varphi$ and (2) for all $\mathfrak{M}'$ such that $\mathfrak{M}\mathcal{R}\mathfrak{M}'$, $\mathfrak{M}' \models \varphi$.

By (1) and the assumption that $\varphi \models_{\mathcal{M}} \psi$ (for some unspecified $\mathcal{M}$ containing all models under consideration in this proof), we conclude that $\mathfrak{M} \models \psi$. Now, ψ is supposed to be $\mathcal{R}$-fragile (over $\mathcal{M}$), so there is some $\mathfrak{M}^*$ $\mathcal{R}$-related to $\mathfrak{M}$ for which $\mathfrak{M}^* \not\models \psi$. In that case, since $\varphi \models_{\mathcal{M}} \psi$, we have $\mathfrak{M}^* \not\models \phi$. But, given (1), this contradicts (2), and this contradiction shows that we could not have φ implying ψ with φ $\mathcal{R}$-stable and ψ $\mathcal{R}$-fragile after all.

To see us now see how this result differs from Restall and Russell's, recall that the latters' Barrier Construction Theorem – our formulation of which appeared as Proposition 3.2 – addresses the consequences of sets of sentences rather than of individual sentences, so Woods and Maguire were hoping for a simplified version of that result which did not use preservation, in their way of setting things out, what Woods and Maguire should have gone for a proof of was this (taking some $\mathcal{M}$ for granted in the background, with implication understood as $\models_{\mathcal{M}}$):

"A set of $\mathcal{R}$-stable sentences does not imply any sentence that is $\mathcal{R}$-fragile,"

or perhaps this with the additional qualifier 'satisfiable' (or 'consistent') on the set of $\mathcal{R}$-stable sentences. But we can easily give a 'disjunctive syllogism' counterexample to this, remembering that $\mathcal{R}$-stable simply means *not* $\mathcal{R}$-fragile; of course for a concrete counterexample, it will help to supply a definite choice of $\mathcal{R}$, so let this be the substructure relation. For this choice of $\mathcal{R}$, $\mathcal{R}$-fragility corresponds to Restall-Russell universality – any model verifying a sentence with this property can be extended to a model not verifying it. As we recall from Restall and Russell's discussion $Fa \vee \forall x(Gx)$ is not fragile with respect to this relation (and not preserved by it either, as they also remarked), since a model verifying the first disjunct cannot be extended to one which does not verify that disjunct however many new object you add (and keep outside of the extension of G). Thus $Fa \vee \forall x(Gx)$ is $\mathcal{R}$-stable, as is $\neg Fa$; this pair of sentence is consistent/satisfiable. But together they imply $\forall x(Gx)$, which is $\mathcal{R}$-fragile, contrary to the would-be theorem. (Thus by the correctness of the Woods–Maguire result, not for arbitrary n, but for the $n = 1$ case of "For any $\mathcal{M}$ and $\mathcal{R} \subseteq \mathcal{M} \times \mathcal{M}$, if $\varphi_1, \ldots, \varphi_n \models_{\mathcal{M}} \psi$, then we cannot have all the φ_i $\mathcal{R}$-stable and ψ $\mathcal{R}$-fragile," one sees that the conjunction of two $\mathcal{R}$-stable formulas is not in general $\mathcal{R}$-stable: as a counterexample take the conjunction of the two formulas just in play: $Fa \vee \forall x(Gx)$ and $\neg Fa$.)

Indeed we knew *a priori* – which, after all, originally meant "according to Prior" – that we could not have a class of sentences Σ to which some sentence and its negation both belong and such that whenever Σ_0 is a consistent subset of Σ with B as a consequence, $B \in \Sigma$, without Σ being the class of all sentences. (At least we have this subject to very weak assumptions about the existence of independent sentences, as detailed in Proposition 1.1.) This is the same reason that Russell [2011] gets only the result mentioned at the end of note 71 and not an unconditional barrier theorem in the style of Restall and Russell [2010]. Woods and Maguire go for a dichotomous classification by starting with a fragility notion for the 'conclusion class' and taking its complement, stability, for the premise class, respectively, as remarked, $\forall\exists$ and $\exists\forall$ notions, whereas Russell's 'premise class' comprises the sentences that are preserved by a context-shift relation ('constant sentences': an $\forall\forall$ notion) and takes its complement ('indexical sentences': an $\exists\exists$ notion) as the conclusion class – again a two-block partition and so by Prior's observation, no straight barrier thesis to be had.

We return to one aspect of Woods and Maguire's formal discussion in the following paragraph, here noting that Woods and Maguire, although [Woods and Maguire, 2017] does not quite convey them accurately, do not contest Restall and Russell's technical results, and worrying mainly, as in the case of Singer touched on in Section 2, that the fragility notions in play – our discussion having concentrated on translation ("$\langle\!\rangle$") fragility to the exclusion of what [Restall and Russell, 2010] calls normative extension – cannot be capturing any intuitive idea of ethicality or normativity. The interested reader is invited to look at the first two paragraphs of §3.3 of [Woods and Maguire, 2017] to see the examples intended to illustrate this charge. The authors then turn to the $\Box p \to Op$ issue which exercised Vranas and Schurz, as cited in Example 3.9. Here again the interested (and preferably patient) reader is referred to their take on what the example shows, since the discussion aims to reveal inappropriate verdicts of descriptiveness delivered by the apparatus of [Restall and Russell, 2010], but uses the mischaracterization mentioned in note 76 of what descriptive sentence are according to Restall and Russell (which is not unconnected with the idea, above, of trying to run the basic Restall–Russell proof using stability, i.e., failure of fragility, in place of preservation).

On the subject of stability, it is instructive to pause over the fact that the conjunction of two $\mathcal{R}$-stable sentences need not be $\mathcal{R}$-stable, and that Woods and Maguire's variation on Restall and Russell does not deliver the general multi-premise version of the latters' barrier the-

sis (or Barrier Construction Theorem: Proposition 3.2 above). Brown [2015] rightly makes the 'trichotomy' point: that this does not vindicate the similarly general version of Hume's Law – [Restall and Russell, 2010]'s Corollaries 24 and 25 – because "Prior is concerned with arguments from the nonmoral to the moral (...) where these are assumed to be exhaustive categories" (p. 3). But he also makes the useful observation that there is nothing like Prior's argument which would make corresponding difficulty for a restricted version of Hume's Law – in a genuinely dichotomous form – where the restriction is to single-premise arguments. Note that instead of saying that that we have a dichotomous 'validity' (as opposed to 'soundness') version of Hume's Law applying to single-premise arguments, we can put this by saying in the terminology used in the Postscripts to Sections 1 and 2 that we can extend a basic ethical/nonethical division so that it becomes exhaustive and still satisfies the condition that the class of ethical statements are closed under converse entailment – the convers of the binary relation of entailment between statements. For our purposes, we can state Brown's observation as a comment on the earlier distillation of Prior's argument in the following way:

Proposition 3.10. *Proposition 1.1 becomes false if, keeping the conditions* (1) *and* (2) *there as they are but restricting* (3) *to the case of* $n = 1$.

Proof. We must show that we can find E, F, and $\mathbb{F}$ such that (condition (1)) $F, \neg F \in \mathbb{F}$, and (condition (2)) F is (classically) independent of another sentence E, and also (condition (3^-), say): for any CL-consistent $A \in \mathbb{F}$, if $A \vdash_{\mathsf{CL}} B$ then $B \in \mathbb{F}$, and yet, by contrast with Prop. 1.1, we have $E \notin \mathbb{F}$. No problem: just let E, F be distinct sentence letters (p, q, say) and define $\mathbb{F} = \{A \mid A \nvdash_{\mathsf{CL}} E\}$ – in other words the element of $\mathbb{F}$ are just those formulas that do not by themselves classically imply q. Conditions (1) and (2) are evidently satisfied by the choice of E, F. Checking condition (3^-), suppose for a contradiction that (*i*) $A \in \mathbb{F}$, (*ii*) if $A \vdash_{\mathsf{CL}} B$ but (*iii*) $B \notin \mathbb{F}$. (*i*) means that $A \nvdash_{\mathsf{CL}} E$, and (*iii*) means that $B \vdash_{\mathsf{CL}} E$: but these together clearly contradict (*ii*). Finally, since $E \vdash_{\mathsf{CL}} E$, we do have $E \notin \mathbb{F}$, as desired. □

The proof given here is what might be called a 'proof of concept' demonstration that no amount of piling up of one-premise inferences can achieve the same counter-Humean effect as Prior's argument with the one-premise rule of $\vee$-introduction and the two-premise disjunctive syllogism rule. If we wanted a more realistic way of setting up our

class $\mathbb{F}$ of ('factual' or) non-ethical sentences we would collect all of its intended non-members rather than just one of them, and take $\mathbb{E}$ (let's call it) to be the set of all basic ethical sentences – the recalling the technical project as described in Section 1 of carving up the terrain of the neither-basic-ethical-nor-basic-nonethical – and setting $\mathbb{F}$ to be

$$\{A \mid \textit{ for all } E \in \mathbb{E} : A \nvdash_{\mathsf{CL}} E\}.$$

The above proof of Proposition 3.10 easily adapts to this more realistic setting. (Very little specific to classical logic was used here – basically just the notions of consistency and independence and the relation between them.)

A question is raised by fact that a dichotomous version of Hume's Law not requiring us to trade in validity for soundness, or to make special exception concern vacuous occurrences of expressions, is available when restricted to one-premiss inferences even though it is not available when to such inferences we add those licensed by disjunctive syllogism are permitted. Since classically, disjunctive syllogism is essentially, give or take a double negation equivalence. Modus Ponens for the material conditional, the question arises as to how what has just been said can survive the observation that Modus Ponens (and in fact more than one-premise rules generally) can be replaced by one-premise rules in an axiomatic presentation of a good many logics, classical logic included.[86] Readers for whom this question is of interest will be able to extract an answer from the either of the papers cited in the footnote just flagged, in the case of the first reference by attending to the passage indicated by the ellipsis in the above quotation, and in the second by looking at the discussion of a number of different rules going under the name 'Modus Ponens'.

Returning more directly to Brown's discussion, recall that in Section 2, we raised an eyebrow at the World-Norm Gap thesis from Singer [2015], because according to that thesis if we have $P_1, \ldots, P_n \vdash C$ for a suitable (and indeed, we may suppose, classical) consequence relation $\vdash$) where each P_i is norm-invariant and their conjunction is satisfiable/-consistent, then $P_1 \wedge P_2 \wedge \ldots \wedge P_n \wedge C$ is norm-invariant. The issue was that we wanted to talk about the conclusion of the original argument

[86] Here is how Herrmann and Rautenberg [1990] put this at their p. 334: "As a by-product, we obtain also the remarkable fact that the set T_2 of 2-valued tautologies (...) is axiomatizable by finitely many axioms and unary rules." A simpler proof of this result can be found in Humberstone [2008]; as John Halleck later reminded the author, Porte [1962] had long ago exhibited such an axiomatization (though admittedly one with many more axioms).

C, rather than this new conjunction with all the premises as further conjuncts. However, it is less likely that one would have had this 'someone changed the subject' reaction if we had passed from the original n-premise argument to making a comment on the conjunctive 1-premise argument with the same conclusion C but the new premise $P_1 \wedge \ldots \wedge P_n$. After all, certainly in a classical setting, whether to say the premises together entail a conclusion or instead the conjunction of the premises entails the conclusion – that's not something one would normally lose a lot of sleep over. But as Brown [2015] points out, this gives rise to two readily distinguishable things to mean by Hume's Law, not because the multi-premise and single conjunctive premise arguments differ as to validity, but because whatever criterion of non-ethicality we are using has to be applied in the one case to each of the several premises and in the other to the single conjunctive premise, and thus the arguments may differ in respect of whether they conform to or violate Hume's Law.

By way of explanation as to why it might be plausible to hold, as an account along the above lines must, that two non-ethical statements can have an ethical conjunction, Brown writes [2015, p. 4]:

> To illustrate, consider the property of being offensive. The sentence "You are either a genius or an idiot" is not offensive. Nor is the sentence "You are no genius"; it is compatible with your being of quite respectable intelligence. But the conjunction, "You are either a genius or an idiot, and you are no genius," is offensive. The reason is that the conjunction says something extra, over and above what is said by either conjunct, namely, that you are an idiot. The offensiveness results only from the two conjuncts combining together; it is not present in either on its own.

Brown goes on to point out that a conjunction, one conjunct of which is offensive, is itself offensive, whatever the other conjunct may be. In that respect as well as in those evident in the above passage, offensiveness is like ethicality on a commitment based view – the kind of view discussed especially in the Postscript to Section 2. We should note, though that there are several ways of giving offence, and apart from being offensive by being insulting, as in the above passage, there is the use of language found offensive by an addressee – for example by swearing, of this or that kind. A disjunction in which one disjunct is offensive in any such way is itself offensive, and one can imagine someone thinking of this as a the more appropriate parallel. The ethicality of a component would infect any compound containing it.

Indeed we do not have to imagine such a position, we can read about it in print: in their very different ways Beall [2014] and Fine [2018] make

suggestions of this kind. In Beall's case the idea, perhaps proposed somewhat facetiously, is that one use the three-valued truth-tables associated with Dmitri Bochvar and known also under the rubric 'Weak Kleene', in which classically behaving truth and falsity is joined by a third value that infects any compound once a component has that value – and that third value will serve as a marker for ethicality. The new value is undesignated, though, which may suits the moral nihilist but is out of place in a more neutral response to Prior's criticism of Hume's Law. For this reason, Beall also considers another option, at least for ethicality with a deontic source: Kripke style models with world-relative truth in the Bochvar three-valued scheme OA being true at a world when all accessible worlds have A true, OA false[87] if all accessible worlds have A false, and OA taking the infectious third value in all other cases – these other cases now including cases in which A takes one or other of the 'classical' values at all accessible worlds, but not uniformly so.[88] It is not clear why we should be forced to give up the obligatory/permissible distinction, though.

Another option might have been to take the infectious value as designated (*à la* [Ciuni and Carrara, 2016]), but this is just as inappropriately unselective as the first option, now looking favourably rather than unfavourably on all moral judgments at once. If anything finitely many-valued might be appropriate in this area, perhaps it a variation on the direct product of the of the two-valued Boolean matrix, whose elements we may call T and F with the two-valued Bochvar matrix, whose elements we may call e and $\bar{e}$ for *ethical* and *non-ethical*, the former being the infectious element. Thus the values $\langle T, e\rangle$ and $\langle T, \bar{e}\rangle$ for the ethical and non-ethical truths, resp., and $\langle F, e\rangle$ and $\langle F, \bar{e}\rangle$ for the ethical and non-ethical falsehoods, the designated values being the former pair (which is why we do not here have a traditional product matrix, which would require for designating that the first and second entries in a designated pair be designated in their respect factor matrices). All the second coordinate is doing here is keeping track of infectious ethicality; which is not to say that this, or the previous suggestion, would suit Beall's purposes, since they do not result in invalidating Prior's $\vee$-introduction inference. The 'infectious' theme we saw also with Fine's proposal in [Fine, 2018], described above (note 61 and the text to which it is appended) in the terminology *dominant* as opposed to *recessive*. Again there is

[87] More precisely: OA having a true negation, since any undesignated value is essentially a species of falsity – as Dummett, Suszko, and Scott have variously observed.

[88] Observe that this recipe for assigning values to OA is only consistent if every world as at least one world accessible to it – an assumption of standard deontic logic.

no intention to invalidate $\vee$-introduction: Fine's way with Hume's Law (which he considers only in connection with one-premise inferences, or entailments, and, it will be recalled, at the level of – albeit structured – propositions rather than of the sentences that express them) is that although the dominance of the normative in the construction of propositions gives us cases of a descriptive proposition entailing a normative proposition, in such cases the former entails suitably de-normativized core of the latter. Successive formulations (and Fine progresses through five of these) of the resulting Humean principle further tweak the way the de-normativized core is characterized. A representative intermediate case, the third approximation to the final proposal (the latter involving too many concepts to explain here) is given as:

(***) No descriptive proposition P entails a normative proposition Q unless P entails $(Q)^{\mathrm{D}}$ or Q is necessary.

Here $(Q)^{\mathrm{D}}$ is the current incarnation of the de-normativized core of Q and is defined to be $Q \cap D$ where D is the set of descriptive states (see note 61 and adjacent text above).[89] (***) is reminiscent at the propositional level of what at the sentential level would be a kind of interpolation theorem, specifically one promising "left uniform interpolants" (because $(Q)^{\mathrm{D}}$ is chosen independently of P, the latter being in the 'left-hand' – or *premise* or *antecedent* position: for a careful definition and relevant references, see the opening paragraph of van Gool et al. [van Gool *et al.*, 2017]). The analogy is only approximate, since $(Q)^{\mathrm{D}}$ may contain descriptive material absent from P. Whether the policy of 'normative infection' is pursed sententially or propositionally in the case of the familiar basic sentence connectives (or the corresponding propositional constructions), this will surely have to stop somewhere if deactivating – or 'protective,' as it is put in note 31 – contexts are on the linguistic menu, on pain of conflating the two notions of the ethical distinguished in that note: ethicality as potentially expressive of an ethical stance vs. ethicality as involving the deployment of ethical concepts.

Section Postscript (ii): De-universalizing Aboutness Two examples of the syntactic process we called de-universalzing toward the end of this section are mentioned in Humberstone [1996], the first only a suggestive analogy to introduce the second, and both of them associated

[89] Fine calls $(Q)^{\mathrm{D}}$ the *disjunct descriptive content* of Q, and for a subsequent honing of the principle we are introduced to the *conjunct descriptive content* $(Q)_{\mathrm{D}}$ of Q, which throws out the normative components from all the consistent truthmakers for Q.

with the work of David Lewis. For the first, consider an initial characterization by Lewis, with which he was not completely satisfied (because of its uninformative potential circularity rather than its incorrectness), of *intrinsic* properties as properties w.r.t. which any two (qualitative) duplicates agree – either both or neither having the property in question.[90] The "any two" here marks an $\forall\forall$ prefix, and removing the first $\forall$ gives for any property P a property of having P intrinsically: x has this new property just in case for all y, if x and y are duplicates, y has the P. Since x is in the relevant sense a duplicate of itself, having the property P intrinsically does imply having the property P, but it does not imply that P is itself an intrinsic property, since there may be other pairs of individuals which are duplicates but do not agree w.r.t. P. Thus supposing that *being circular* is an intrinsic property but *being within a metre of something circular* is not, a circular ring still has the latter property intrinsically since any duplicate of it will agree with it w.r.t. the property of being within a metre of something circular. On the other hand, an iron nail sitting next to such a ring does not have the property *being within a metre of something circular* intrinsically.

That example of de-universalizing serves as a warm-up exercise for the case of a statement's being entirely about a subject matter, as this is conceived in Lewis [1988]. A subject matter here is thought of as a partition of the set of worlds: those that are alike in respect of that subject matter. If M is a subject matter then we denote the corresponding equivalence relation by $\equiv_M$.[91] A statement S is *entirely about* a subject matter M just in case for any worlds w, w', if $w \equiv_M w'$ then S is true at w iff S is true at w'. De-universalizing, we get the following the following property of a world w: being such that for all w', if $w \equiv_M w'$ then S is true at w iff S is true at w'. In other words, at w the truth-value of S is settled by the subject matter M. Just as a property possessed intrinsically by an object need not be an intrinsic property, so such an M-settled statement need not a statement entirely about M. One of the subject matters mentioned by Lewis in [1988] is that of the seventeenth century and another, that of the eighteenth century, with associated equivalence relations of exact match of worlds over the respective time

[90] The details of Lewis's various attempts at throwing light on this topic, together with all the relevant references, can be found in Marshall and Weatherson [2018].

[91] This will suit our purposes here, though as one of several refinements of Lewis's account, Yablo [2014, p. 36] suggests we don't actually want partitions and equivalence relations here, since transitivity will fail for such subject-matters as *approximately how many stars there are*, even when the vagueness of "approximately" is removed (e.g., being replaced by "to within 100").

periods. The statement:

(†) There were dinosaurs in Europe in the seventeenth century but they were all extinct by the end of the eighteenth century.

is not entirely about the seventeenth century, since worlds could be relevantly equivalent in respect of their seventeenth centuries but differ in respect of whether that statements was true in them: in one dinosaurs are extinct by 1710, and in the other, not until 2010, say. But in the actual world the truth-value of (†) is settled by the 17th century subject matter, since equivalence w.r.t. that subject matter suffices for the falsity of the conjunction in any 17th-century matching world. In this case the statement is settled as false by the subject matter, or M^--settled in the actual world, as it is put in [Humberstone, 1996], where M is the subject matter in question, as opposed to M^+-settled, or M-settled as true in w.

We can think of taking the property of being M^+-settled into the object language as a modal operator, $\Box$ with accessibility relation M, which we may write as $\Box_M$. Thus $\Box_M A$ is true at w when A is M^+-settled in w. Indeed, such an operator was suggested in Yablo [2014], pp. 32–34, though the focus there is rather more strongly on the dual operator $\Diamond_M$, with Yablo's preferred reading of $\Diamond_M A$ being something along the lines of A's being true about M at the world in question, acknowledging that this may not be of much interest if A says nothing about M – so perhaps a safer reading would be in terms of A's not being false about M in w, i.e., A's not saying anything false about M at w. (Yablo actually writes "m" rather than "M".)

If, as in Humberstone [1996], one wants to use this kind of machinery to discuss statements with a Gibbard–Karmo–Singer semantics in mind, then since for the truth-evaluation of a sentence, one needs not only a world a but also an ethical standard (to use Karmo's term), the subject-matters should be partitions of the set of 'Gibbard-worlds': Singer's $\langle w, n\rangle$ pairs. And here two especially salient subject-matters called in [Humberstone, 1996] M_{nat} and M_{eth} force themselves on one's attention (the subscripts suggesting 'natural(istic)' and 'ethical' respectively – though in [Humberstone, 1996] '*eth*' appeared as '*eval*'). The associated equivalence relations $\equiv_{nat}$ and $\equiv_{eth}$ relates any $\langle w, n\rangle$ to $\langle w', n'\rangle$ if and only if, for the former $w = w'$, and, for the latter, when $n = n'$. As a fair approximation, the basic ethical and basic non-ethical statements can be taken as those entirely about M_{eth} and those entirely about M_{nat}, respectively (though since anything true at every or false at every $\langle w, n\rangle$ pair will then count as both, contrary to our expectation to

have these classes of statements disjoint). And, following Lewis's lead in [Lewis, 1988], we note that whereas the statements entirely about M_{eth} are closed under negation as are those entirely about M_{nat}, and indeed those entirely about any given subject matter, and not just closed under negation but under all Boolean operations, though not under entailment or under converse entailment. If we want to get classes of statements which are closed under converse entailment, we can do so by replacing "is entirely about M" with "entails something contingent which is entirely about M," giving essentially one of Lewis's glosses on "partly about M"[92] – though now we lose the property of being closed under negation. Recall the first passage from Karmo quoted in the Postscript to Section 2, noting the tension between these features. (Lewis [1988] and Karmo [1988] both appeared in the same year: 1988.) The notion of ethicality as entailing something entirely about M_{eth} is probably the simplest such notion embodying the 'commitment' idea in play in that Postscript, though the alternative to Karmo's suggestion in the quoted passage was, like his own, a world-relative notion. In Karmo's case, translated into the present concepts, non-ethicality at w is a matter of being M_{nat}-settled at w and thus, ethicality at w is a matter of not being thus settled. One could equivalently say (non-)ethicality at $\langle w, n\rangle$ here, since the this does not depend on any particular choice of n. On the other hand, ethicality at w according to the 'enthymematic' proposal of Humberstone [1982a] is a world-relativized variant on being partly about M_{eth}, but instead of being a matter of entailing something contingent entirely about M_{eth}, is a matter of being such that it together with additional premises true at w and entirely about M_{nat}, entails something contingent which is entirely about M_{eth}. This gloss on [Humberstone, 1982a] is taken from Humberstone [1996], in which further details on the relations between M_{eth} (or "M_{eval}") and M_{nat} are related. The imaginary interlocutor summoned up by Geach in the passage quoted in Example

[92]The gloss in question is what Lewis calls the part-of-content notion of partial aboutness, though he does not seem to do the equivalent in his negative way of describing it (the content of a statement being the set of worlds at which it is false) of inserting the word *contingent*, as here: since with the classical assumptions in force here and in Lewis [1988], every statement entails any logical truth and that is entirely about every subject matter, we need to exclude such cases when we say "entails a statement entirely about M" if it is not to apply across the board to all statements. Not that *contingent* is really the right word, in the first place because here we only need to exclude necessary truths rather than all non-contingent statements, a and secondly because even in making that adjust we are in the wrong modality, it being logical truths (true at all $\langle w, n\rangle$ pairs) rather than necessary truths (true at all w) that need to be excluded.

2.5, with its reference to the supplementary non-ethical premise that one could reach for in the envisaged disjunctive syllogism step there, takes very much the line developed in [Humberstone, 1982a] – written, as it happens, without knowledge of Geach [1976].

References

[Anderson and Belnap, 1975] A. R. Anderson and N. D. Belnap, *Entailment: the Logic of Relevance and Necessity, Vol. I*, Princeton University Press, Princeton, NJ 1975.

[Atkinson, 1958] R. F. Atkinson, 'The Autonomy of Morals', *Analysis* **18** (1958), 57–62.

[Basl and Coons, 2017] John Basl and Christian Coons, 'Ought to Is: the Puzzle of Moral Science', pp. 160–186 in Russ Shafer-Landau (ed.), *Oxford Studies in Metaethics* **12**, Oxford University Press, Oxford 2017.

[Beall, 2014] J. C. Beall, 'A Neglected Reply to Prior's Dilemma', pp. 203–208 in J. Maclaurin (ed.), *Rationis Defensor: Essays in Honour of Colin Cheyne*, Springer 2014.

[Borowski, 1976] E. J. Borowski, 'A Pyrrhic Defence of Moral Autonomy', *Philosophy* **52** (1976), 455–466.

[Borowski, 1980] E. J. Borowski, 'Moral Autonomy Fights Back', *Philosophy* **55** (1980), 95–100.

[Brown, 2014] Campbell Brown, 'Minding the Is-Ought Gap', *Journal of Philosophical Logic* **43** (2014), 53–69.

[Brown, 2015] Campbell Brown, 'Two Versions of Hume's Law,' *Journal of Ethics and Social Philosophy* **9** (2015), 7pp. in Issue 1 (online).

[Büring, 2007] Daniel Büring, 'Semantics, Intonation, and Information Structure', pp. 445–473 in G. Ramchand and C. Reiss (eds.), *The Oxford Handbook of Linguistic Interfaces*, Oxford University Press, Oxford 2007.

[Ciardelli and Roelofsen, 2017] Ivano Ciardelli and Floris Roelofsen, 'Hurford's Constraint, the Semantics of Disjunction, and the Nature of Alternatives', *Natural Language Semantics* **25** (2017), 199–222.

[Ciuni and Carrara, 2016] Roberto Ciuni and Massimiliano Carrara, 'Characterizing Logical Consequence in Paraconsistent Weak Kleene', pp. 165–176 in L. Felline, A. Ledda, F. Paoli, and E. Rossanese (eds.), *New Directions in Logic and the Philosophy of Science*, College Publications, London 2016.

[Cueno and Shafer-Landau, 2014] T. Cuneo and R. Shafer-Landau, 'Moral Fixed Points: New Directions for Moral Nonnaturalism', *Philosophical Studies* **171** (2014), 399–443.

[Diller, 2003] Antoni Diller, 'Retransmittability and Empirical Propositions', pp. 243–247 in F. H. van Eemeren, J. A. Blair, C. A. Willard and A. F. Snoek Henkemans (eds.), *Procs. of the Fifth Conference of the International Society for the Study of Argumentation*, SicSat Amsterdam 2003.

[Dreier, 2002] James Dreier, 'Meta-Ethics and Normative Commitment', *Philosophical Issues* **12** (2002), 241–263.

[Dummett, 1978] M. A. Dummett, 'The Justification of Deduction', pp. 290–318 in *Truth and Other Enigmas*, Duckworth, London 1978.

[Dworkin, 2012] Ronald Dworkin, *Justice for Hedgehogs*, Harvard University Press, Cambridge, Mass. 2012.

[Elgin, 2013] Catherine Z. Elgin, 'Fact-Value Distinction', pp. 1881–1887 in H. LaFollette (ed.), *The International Encyclopedia of Ethics*, Blackwell Publishing, Oxford 2013.

[Fine, 2017] Kit Fine, 'Truthmaker Semantics', pp. 556–577 in Bob Hale, C. Wright and A. Miller (eds.), *A Companion to the Philosophy of Language*, Second edn., *Volume 2*, Wiley-Blackwell, Oxford 2017.

[Fine, 2018] Kit Fine, 'Truthmaking and the Is-Ought Gap' online first at *Synthese* (as of 2018) https://doi.org/10.1007/s11229-018-01996-8

[Gautam, 1057] N. D. Gautam, 'The Validity of Equations of Complex Algebras', *Archiv. Math. Logik Grundlagenforschung* **3** (1957), 117–124.

[Geach, 1976] P. T. Geach, 'Morally Significant Theses', *Open Mind* **4** (1976), 5–12.

[Geach, 1976a] P. T. Geach, 'Murder and Sodomy', Philosophy **51** (1976a), 473–476.

[Geach, 1977] P. T. Geach, 'Again the Logic of "Ought" ' *Philosophy* 52 (1977), 473–476.

[Geach, 1979] P. T. Geach, 'Kinds of Statement', pp. 221–235 of C. Diamond and J. Teichman (eds.), *Intention and Intentionally: Essays in Honour of G. E. M. Anscombe*, Harvester Press, Brighton 1979.

[Geach, 1982] P. T. Geach, 'Moral Autonomy Still Refuted', *Philosophy* 57 (1982), 127–129.

[Gewirth, 1979] Alan Gewirth, 'On Deriving a Morally Significant "Ought"', *Philosophy* **54** (1979), 231–232.

[Garcia, 1995] J. L. A. Garcia, 'Are "Is" to "Ought" Deductions Fallacious?: On a Humean Formal Argument', *Argumentation* **9** (1995), 543–552.

[Gibbard, 1990] Allan Gibbard, *Wise Choices, Apt Feelings*, Harvard University Press, Cambridge, Mass. 1990.

[Gibbard, 2003] Allan Gibbard, *Thinking How to Live*, Harvard University Press, Cambridge, Mass. 2003.

[Gibbard, 2005] Allan Gibbard, 'Truth and Correct Belief', *Philosophical Issues* **15** (2005), 338–350.

[Gibbard, 2012] Allan Gibbard, *Meaning and Normativity*, Oxford University Press, Oxford 2012.

[van Gool *et al.*, 2017] Samuel J. van Gool, George Metcalfe, Constantine Tsinakis, 'Uniform Interpolation and Compact Congruences', *Annals of Pure and Applied Logic* **168** (2017) 1927–1948.

[Greenspan, 1975] Patricia Greenspan, 'Conditional Oughts and Hypothetical

Imperatives', *Journal of Philosophy* **72** (1975), 259–276.

[Guevara, 2008] Daniel Guevara, 'Rebutting Formally Valid Counterexamples to the Humean "is-ought" Dictum', *Synthese* **164** (2008), 45–60.

[Hare, 1977] R. M. Hare, 'Geach on Murder and Sodomy', *Philosophy* **52** (1977), 467–472.

[Heathcote, 2010] Adrian Hathcote, 'Hume's Master Argument', pp. 92–117 in Pigden [Pigden, 2010].

[Herrmann and Rautenberg, 1990] B. Herrmann and W. Rautenberg, 'Axiomatization of the De Morgan Type Rules', *Studia Logica* **49** (1990), 333–343.

[Hill, 2008] Scott Hill, ' "Is"–"Ought" Derivations and Ethical Taxonomies', *Philosophia* **36** (2008), 545–566.

[Hill, 2009] Scott Hill, 'Good News for the Logical Autonomy of Ethics', *Argumentation* **23** (2009), 277–283

[Horwich, 2018] Paul Horwich, 'Is TRUTH a normative concept?', *Synthese* **195** (2018), 1127–1138.

[Humberstone, 1982] Lloyd Humberstone, 'Necessary Conclusions', *Philosophical Studies* **41** (1982), 321–335.

[Humberstone, 1982a] Lloyd Humberstone, 'First Steps in Philosophical Taxonomy', *Canadian Journal of Philosophy* **12** (1982), 467–478.

[Humberstone, 1983] Lloyd Humberstone, 'The Background of Circumstances', *Pacific Philosophical Quarterly* **64** (1983), 19–34.

[Humberstone, 1995] Lloyd Humberstone, Review of J. P. Cleave, *A Study of Logics*, and A. Koslow, *A Structuralist Theory of Logic*, *Australasian Journal of Philosophy* **73** (1995), 475–481.

[Humberstone, 1996] Lloyd Humberstone, 'A Study in Philosophical Taxonomy', *Philosophical Studies* **83** (1996), 121–169.

[Humberstone, 1997] Lloyd Humberstone, 'Two Types of Circularity', *Philosophy and Phenomenological Research* **58** (1997), 249–281.

[Humberstone, 2000] Lloyd Humberstone, 'What *Fa* Says About *a*', *Dialectica* **54** (2000), 1–28.

[Humberstone, 2008] Lloyd Humberstone, 'Replacing Modus Ponens With One-Premiss Rules', *Logic Journal of IGPL* **16** (2008), 431–451.

[Humberstone, 2013] Lloyd Humberstone, 'Inverse Images of Box Formulas in Modal Logic', *Studia Logica* **101** (2013), 1031–1060.

[Humberstone, 2014] Lloyd Humberstone, 'Power Matrices and Dunn-Belnap Semantics: Reflections on a Remark of Graham Priest', *Australasian Journal of Logic* **11** (2014), 14–45.

[Humberstone, 2016] Lloyd Humberstone, *Philosophical Applications of Modal Logic*, College Publications, London 2016.

[Humberstone, 2019] Lloyd Humberstone, 'Supervenience, Dependence, Disjunction', *Logic and Logical Philosophy* **28** (2019), 3–135.

[Humberstone, 2020] Lloyd Humberstone, 'Explicating Logical Independence', *Journal of Philosophical Logic*, **49** (2020), 135–218.

[Hurka, 1980] Thomas Hurka, 'Geach on Deriving Categorical "Oughts" ', *Philosophy* **55** (1980), 101–104.

[Jackson, 1974] Frank Jackson, 'Defining the Autonomy of Ethics', *Philosophical Review* **83** (1974), 88–96.

[Jackson, 1985] Frank Jackson, 'On the Semantics and Logic of Obligation', *Mind* **94** (1985), 177–195.

[Jackson, 1987] Frank Jackson, *Conditionals*, Basil Blackwell, Oxford 1987.

[Jackson, 2013] Frank Jackson, 'Autonomy of Ethics', pp. 459–465 in H. LaFollette (ed.), *The International Encyclopedia of Ethics*, Blackwell Publishing, Oxford 2013.

[Jaggar, 1974] Alison Jaggar, 'It Does Not Matter Whether We Can Derive "Ought" from "Is"', *Canadian Journal of Philosophy* **3** (1974), 373–379.

[Jones and Pörn, 1986] Andrew J. I. Jones and Ingmar Pörn, '"Ought" and "Must"', *Synthese* **66** (1986), 89–93.

[Karmo, 1988] Toomas Karmo 'Some Valid (but no Sound) Arguments Trivially Span the "Is"–"Ought" Gap', *Mind* **97** (1988), 252–257.

[Kolodny and MacFarlane, 2010] Niko Kolodny and John MacFarlane, 'Ifs and Oughts', *Journal of Philosophy* **107** (2010), 115–143.

[Lemmon, 1965] E. J. Lemmon, *Beginning Logic*, Nelson, London 1965.

[Lewis, 1988] David Lewis, 'Statements Partly about Observation', *Philosophical Papers* **17** (1988), 1–31.

[Maguire, 2015] Barry Maguire, 'Grounding the Autonomy of Ethics', pp. 188–215 in R. Shafer-Landau (ed.), *Oxford Studies in Metaethics, Vol. 10*, Oxford University Press, Oxford 2015.

[Maitzen, 1998] Stephen Maitzen, 'Closing the "Is"–"Ought" Gap', *Canadian Journal of Philosophy* **28** (1998), 349-366.

[Maitzen, 2008] Stephen Maitzen, 'Anti-Autonomism Defended: A Reply to Hill', *Argumentation* **36** (2008), 567–574.

[Maitzen, 2010] Stephen Maitzen, 'Moral Conclusions from Non-moral Premises', pp. 290–309 in Pigden [Pigden, 2010].

[Makinson, 1999] David Makinson, 'On a Fundamental Problem of Deontic Logic', pp. 29–53 in P. McNamara and H. Prakken (eds.) *Norms, Logics and Information Systems: New Studies in Deontic Logic and Computer Science*, IOS Press, Amsterdam 1999.

[Mares, 1992] E. D. Mares, 'Andersonian Deontic Logic', *Theoria* **58** (1992), 3–20.

[Mares, 2010] E. D. Mares, 'Supervenience and the Autonomy of Ethics: Yet Another Way in which Relevant Logic is Superior to Classical Logic', pp. 272–289 Pigden [Pigden, 2010].

[Marshall and Weathersn, 2018] Dan Marshall and Brian Weatherson, 'Intrinsic vs. Extrinsic Properties', Edward N. Zalta (ed.), *Stanford Encyclopedia of Philosophy* (Spring 2018 Edition), Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/spr2018/entries/intrinsic-

extrinsic/>.

[Mavrodes, 1964] George Mavrodes, ' "Is" and "Ought" ', *Analysis* **25** (1964), 42–44.

[Mavrodes, 1968] George Mavrodes, 'On Deriving the Normative from the Non-Normative', in *Papers of the Michigan Academy of Arts and Sciences* **53** (1968), 353–365.

[Morgan, 1973] Charles G. Morgan, 'Drawing Dichotomies via Formal Languages', *Southern Journal of Philosophy* **11** (1973), 216–227.

[Morscher, 2016] Edgar Morscher, 'The Descriptive-Normative Dichotomy and the So Called Naturalistic Fallacy', *Analyse & Kritik* **38** (2016), 317–337.

[Nelson, 1995] Mark T. Nelson, 'Is it Always Fallacious to Derive Values from Facts?', *Argumentation* **9** (1995), 553–562.

[Nelson, 2003] Mark T. Nelson, 'Who Needs Valid Moral Arguments?' *Argumentation* **17** (2003), 35–42.

[Nelson, 2007] Mark T. Nelson, 'More Bad News for the Logical Autonomy of Ethics', *Canadian Journal of Philosophy* **37** (2007), 203–216.

[Pigden, 1989] Charles R. Pigden, 'Logic and the Autonomy of Ethics', *Australasian Journal of Philosophy* **67** (1989), 127–151.

[Pigden, 2007] Charles R. Pigden, 'Nihilism, Nietzsche and the Doppelganger Problem', *Ethical Theory and Moral Practice* **10** (2007), 441–456.

[Pigden, 2010] Charles R. Pigden (ed.), *Hume on 'Is' and 'Ought'*, Palgrave Macmillan, NY 2010.

[Pigden, 2010] Charles R. Pigden, 'On the Triviality of Hume's Law: A Reply to Gerhard Schurz', pp. 217–238 in [Pigden, 2010].

[Pigden, 2013] Charles R. Pigden, 'Is–Ought Gap', pp. 2793–2801 in H. LaFollette (ed.), *The International Encyclopedia of Ethics*, Blackwell Publishing, Oxford 2013.

[Pigden, 2016] Charles R. Pigden, 'Hume On Is and Ought: Logic, Promises and the Duke of Wellington', pp. 401–415 in Paul Russell (ed.), *Oxford Handbook on David Hume*, Oxford University Press, Oxford 2016.

[Porte, 1962] Jean Porte, 'Un Système Logistique Très Faible pour le Calcul Propositionnel Classique', *Comptes Rendus Hebdomadaires des Sééances de l' Académie des Sciences* **254** (1962), 2500–2502.

[Potts, 2005] Christopher Potts, *The Logic of Conventional Implicatures*, Oxford University Press, Oxford 2005.

[Prior, 1960] A. N. Prior, 'The Autonomy of Ethics', *Australasian Journal of Philosophy*, **38** (1960), 199–206. (Reprinted as Chapter 10 in P. T. Geach and A. J. Kenny (eds.), *A. N. Prior: Papers on Logic and Ethics*, Duckworth, London 1976, and as Chapter 1.1 in Pigden [Pigden, 2010].)

[Prior, 1962] A. N. Prior, *Formal Logic* (Second edn.), Oxford 1962. (First edn. 1955.)

[Remnant, 1959] Peter Remnant, 'Professor Rynin on the Autonomy of Morals', *Mind* **68**, (1959), 252–255.

[Rescher, 1990] Nicholas Rescher, 'How Wide is the Gap Between Fact and Value?', *Philosophy and Phenomenological Research* **50** Supplement (1990), 297–319.

[Restall and Russell, 2010] Greg Restall and Gillian Russell, 'Barriers to Implication' pp. 243–259 in Pigden [Pigden, 2010].

[Routley and Routley, 1969] R. Routley and V. Routley, 'A Fallacy of Modality', *Noûs* **3** (1969), 129–153.

[Russell, 2010] Gillian Russell, 'In Defence of Hume's Law', pp. 151–161 in Pigden [Pigden, 2010].

[Russell, 2011] Gillian Russell, 'Indexicals, Context-sensitivity and the Failure of Implication', *Synthese* **183** (2011), 143–160.

[Rynin, 1957] David Rynin, 'The Autonomy of Morals', *Mind* **66** (1957), 308–317.

[Sint Croix and Thomason, 2014] Catharine Saint Croix and Richmond H. Thomason, 'Chisholm's Paradox and Conditional Oughts', pp. 192–207 in F. Cariani, D. Grossi, J. Meheus, X. Parent (eds.), *Deontic Logic and Normative Systems* (12th International Conference, DEON 2014, Ghent, Belgium, July 12-15), LNAI 8554, Springer, Cham, Switzerland 2014.

[Schotch and Jennings, 1981] P. K. Schotch and R. E. Jennings, 'Non-Kripkean Deontic Logic', pp. 149–162 in R. Hilpinen (ed.), *New Studies in Deontic Logic*, Reidel, Dordrecht 1981.

[Schurz, 1994] Gerhard Schurz, 'Hume's Is-Ought Thesis in Logics with Alethic-Deontic Bridge Principles', *Logique et Analyse* **37** (1994), 265–293.

[Schurz, 1997] Gerhard Schurz, *The Is-Ought Problem: An Investigation in Philosophical Logic*, Kluwer, Dordrecht 1997.

[Schurz, 2010] Gerhard Schurz, 'Non-trivial Versions of Hume's Is-Ought Thesis', pp. 198–216 in Pigden [Pigden, 2010]

[Schurz, 2010a] Gerhard Schurz, Comments on Restall and Russell [Restall and Russell, 2010], pp. 268–271 in [Pigden, 2010].

[Scott, 1974] Dana Scott, 'Completeness and Axiomatizability in Many-Valued Logic', pp.188–197 in L. Henkin *et al.* (eds.), *Procs. of the Tarski Symposium*, American Mathematical Society, Providence, Rhode Island 1974.

[Searle, 1969] John Searle, 'How to Derive "Ought" from "Is" ', *Philosophical Review* **73** (1964), 43–58. Reprinted in pp. 120–134 in W. D. Hudson (ed.), *The Is/Ought Question*, Macmillan, London 1969.

[Sen, 1966] Amartya K. Sen, 'Hume's Law and Hare's Rule', *Philosophy* **41** (1966), 75–79.

[Shaw, 1965] P. D. Shaw, 'On the Validity of Arguments from Fact to Value-Judgements', *Philosophical Quarterly* **18** (1965), 249–255.

[Shorter, 1961] J. M. Shorter, 'Professor Prior on the Autonomy of Ethics', *Australasian Journal of Philosophy* **39** (1961), 286–287. (Reprinted as Chapter 1.2 in Pigden [Pigden, 2010].)

[Singer, 2015] Daniel Singer, 'Mind the Is-Ought Gap', *Journal of Philosophy*

112 (2015), 193–210.

[Singer, 1973] Peter Singer, 'The Triviality of the Debate over "Is–Ought" and the Definition of "Moral" '*American Philosophical Quarterly* **10** (1973) 51–56.

[Sinnott-Armstrong, 2000] Walter Sinnott-Armstrong, 'From "Is" to "Ought" in Moral Epistemology', *Argumentation* **14** (2000), 159–174.

[Sinnott-Armstrong, 2006] Walter Sinnott-Armstrong, *Moral Skepticisms*, Oxford University Press, New York 2006.

[Sobel, 2003] J. H. Sobel, 'The Naturalistic Fallacy and Hume's Law', pp. 213–226 in K. Segerberg and R. Sliwinski (eds.), *Logic, Law, Morality: Thirteen Essays in Practical Philosophy in Honour of Lennart Åqvist*, Uppsala Philosophical Studies 51, Uppsala University, 2003.

[Urmson, 1953] J. O. Urmson, 'Some Questions Concerning Validity', *Revue Internationale de Philosophie* **7** (1953), 217–229; reprinted pp. 120–133 in A. Flew (ed.), *Essays in Conceptual Analysis*, Macmillan, London 1956.

[Vranas, 2007] Peter B. M. Vranas, 'I Ought, Therefore I Can', *Philosophical Studies* **136** (2007), 167–216.

[Vranas, 2010] Peter B. M. Vranas, Comments on Restall, Russell and Vranas [Restall and Russell, 2010], pp. 260–267 in [Pigden, 2010].

[Vranas, 2018] Peter B. M. Vranas, ' "Ought" Implies "Can" but Does Not Imply "Must': An Asymmetry between Becoming Infeasible and Becoming Overridden', *Philosophical Review* **127** (2018), 487–514.

[Wolf, 2015] Aaron Wolf, 'Giving Up Hume's Guillotine', *Australasian Journal of Philosophy* **93** (2015), 109–125.

[Woods and Maguire, 2017] Jack Woods and Barry Maguire, 'Model Theory, Hume's Dictum, and the Priority of Ethical Theory', *Ergo* **4** (2017), 419–440.

[Yablo, 2014] Stephen Yablo, *Aboutness*, Princeton University Press, Princeton, NJ 2014.

[Zolin, 2000] E. E. Zolin, 'Embeddings of Propositional Monomodal Logics', *Logic Journal of the IGPL* **8** (2000), 861–882.

Lloyd Humberstone
Monash University
Email: lloyd.humberstone@monash.edu

More Concepts and Problems

3
Logics for Supererogation and Allied Normative Concepts

Paul McNamara

Abstract. Supererogation (roughly, going beyond the call of duty or doing more than one must) is a familiar part of moral consciousness, and it is one member of a rich family of associated concepts that have proved challenging to adequately model collectively in deontic logic, as well as in ethical theory. Much of the work done, especially earlier work, important as it is, was at the cusp of logic and ethical theory, with this early work having only sketches of logical frameworks and no formal semantics. Only a small body of work from the late 1980s forward meets minimal standards one comes to expect in deontic logic. This essay surveys much of that earlier work in the 1960s and 1970s, regimenting and developing that work, and evaluating it, and then it turns to subsequent more sophisticated work, expositing, at times developing, and evaluating that work. The result is an overview of this underdeveloped area, and an invitation to develop it further. It also serves as a case study of how work in deontic logic can be highly relevant to ethical theory.

1 The Traditional Scheme & "Standard Deontic Logic" 160
1.1 The Traditional Scheme 160
1.2 From the Traditional Scheme to Standard Deontic Logic . 163

2 Supererogation, indifference, and Urmson's constraint. 168
2.1 Moral Indifference, the Strong Threefold Classification, and Moral Rigor 168
2.2 Supererogation, indifference, optionality, and the Fivefold Classification 170

This chapter is dedicated to Lennart Aqvist and Roderick Chisholm, who both went beyond the call in reaching out supportively at the very early stages of my career.

I would like to thank the following authors for help with particular matters: Mark Brown, Bill deVries, Lou Goble, Risto Hilpinen, Andrew Jones, Simo Knuttila, Shahid Rashan, Ulla Wessels. I would like to offer special thanks to Sven Ove Hansson and Xavier Parent for very helpful comments on the chapter as a whole.

Handbook of Deontic Logic and Normative Systems, Volume 2

2.3 Urmson, friends of supererogation, and the Traditional Scheme 174

3 The early axiology-based work of Chisholm 176
3.1 Chisholm's "Supererogation and Offence" 177
3.2 Regimentation of Chisholm's scheme and assessment . . . 183

4 Chisholm on the logic of requirement & supererogation and kin 201
4.1 The 1964 account in the "Ethics of Requirement" 203
4.2 The 1974 account in "Practical Reason and the Logic of Requirement" 204
4.3 Application and extension of the framework to supererogation and kin 215
4.4 Comparison with the prior frameworks and some challenges/disruptions. 219
4.5 Evaluation of the REQ framework for supererogation and kin 230

5 Doing Well Enough (DWE) 234
5.1 Indifference and optionality 235
5.2 *Must* and *ought* 238
5.3 An ignored construction: *the least you can do* 244
5.4 *Doing more (good) than you have to do* 245
5.5 *"You ought to but don't have to"; "you can but ought not"* 246
5.6 Upshot: a cumulative case for logical and semantic framework 247
5.7 DWE operators: personal non-agential readings 253
5.8 Interlude: revisiting the DWE frame structure 256
5.9 Aretaic (agent-evaluative) notions and DWE 260
5.10 Reflections on the DWE framework 268

6 Some other recent work 271
6.1 Wessels' work, and supererogatory holes 271
6.2 Åqvist's systematic frame constants models 279
6.3 Sven Ove Hanson's work and supererogation 289

7 Conclusion 298

Why is supererogation of interest to ethical theory and deontic logic? First, we routinely conceptualize moral exemplars as doing more than they are required to do. This is often their most salient mark. Secondly, the most famous traditional approaches in normative ethical theory (Virtue ethics, Kantian ethics, and Utilitarianism) have had trouble either recognizing the possibility of supererogation or of giving a minimally satisfactory account of it. Thirdly, supererogation is part of a family of concepts that ethical theorists and deontic logicians have often failed to account for, often stumbling around among them, conflating members of distinct pairs with one another. Representing supererogation coherently is hard, and it requires tackling an enriched array of moral concepts and representing their logical relationships carefully enough to generate a coherent framework. This last fact is reflected in the often touted slogan "the traditional deontic scheme must go!" allegedly backed by the claim that supererogation conflicts with the core of deontic logic.

Consider a number of terms of normative appraisal presupposed by common sense morality that ethicists and logicians have been hard-pressed to represent in an integrated conceptual framework:

permissible	significant
impermissible	good
obligatory (required)	bad
omissible (non-obligatory)	praiseworthy
must	blameworthy
can	praise-blame-neutral
can't	action beyond the call (of duty)
ought	more than you had to do
the least you can do	supererogatory
the best one can do	suberogatory
optional	offence
indifferent	

Now consider just these eight concepts: *the obligatory, the least one can do, the best one can do, action beyond the call, the morally optional, the morally indifferent, the morally significant and the permissibly sub optimal*. The traditional framework (a pre-formal fragment of Standard Deontic Logic), employing notions of what is obligatory and permissible, partitions all actions into three mutually exclusive and jointly exhaustive classes: those which are (overridingly) obligatory, those which are (overridingly) impermissible and those which are neither (optional). At most, this scheme can represent exactly two of the eight aforementioned concepts. For from the standpoint of this scheme, *the obligatory* and

the optimal can't be distinguished, yet common sense allows something optimal to fail to be obligatory. Although *the morally optional* can be represented, the supererogatory, one of its proper subclasses, cannot be. Neither can *the supererogatory* be identified with *the morally optimal*, for that which is obligatory can be optimal, but not supererogatory. Similarly, *the morally indifferent*, another proper subclass of *the optional* – one obviously disjoint from *the supererogatory* – can't be represented. Ditto for *the morally significant* and *the permissibly sub optimal*. Finally, *the minimum that morality demands*, a concept generally neglected in the ethical and deontic literature, despite its importance for common sense morality, finds no place in the traditional scheme. Thus, on the face of it, the traditional scheme is radically *incomplete*. It lacks the resources to demarcate an array of concepts of common sense morality.

In deontic logic, there has been very little formal work done on this subject. With a few exceptions, the typical work that has been done is at the intersection of ethical theory and deontic logic, often at best quasi-formal, and when formal, there is rarely any model theory, just axioms or perhaps only a series of definitions cast in some quasi-formal notation, articulating an enrichment of the traditional conceptual scheme. In this chapter, I try to survey some of the landscape of this work done at the cusp of ethical theory and deontic logic, and in a number of cases, I develop the frameworks considerably, and more formally. This seems a necessary step, perhaps providing a shot in the arm for research in this under-explored and underdeveloped area of deontic logic, an area of significance to ethical theory as well, and an area where deontic logic has much to contribute to sharpened ethical theorizing, while at the same time receiving substantial benefit in return. I site one important instance: the neglected difference between *must* and *ought* is highly relevant to ethical theory, but also to practical reason and normative reasoning. Note well that it is *must* not *ought* that is plausibly linked to the *can* and *can't* of permissibility and impermissibility in traditional ways, so that the focus on *ought* in both ethical theory and deontic logic does not have the continuity with traditional concerns with obligation that has been largely presupposed throughout the twentieth century in ethical theory and deontic logic. Yet the distinction between the two is of the first importance in getting clear about the conceptual landscape of supererogation [McNamara, 1994].

Sections 1 and 2 of this essay set the stage and outline some of the conceptual landscape of what I call "the traditional deontic scheme" and some of the intuitive expressive enrichment called for to make a place for supererogation, as well as noting some of the intuitive logical

connections among the enriched set of concepts. Section 3 examines Chisholm's important quasi-formal work in this area. Chisholm stands alone before the 1990s in having made a sustained effort to try to make a place for supererogation and its associated family of concepts. Although his contribution on contrary to duty imperatives [Chisholm, 1963b] is one of the most well-known and oft-cited landmarks in the philosophical literature on deontic logic, his contributions on our chapter topic are under-appreciated attempts to again contribute to deontic logic. In Section 3 we look at one key axiological approache to the conceptual analysis of supererogation and kindred notions, that of the influential (in ethics primarily) and seminal "Supererogation and Offence". We can only mention in passing the related work by Chisholm and Sosa on the logic of intrinsic value and supererogation. In Section 4, we turn to Chisholm's well-known and influential (in deontic logic) logic of requirement and to its much less well-known applications to supererogation and kindred concepts. These three approaches have not, been carefully scrutinized, despite their being influential. We take significant steps in doing so here for the first and third approaches, developing them carefully, albeit still leaving much aside. In Section 5, we turn to McNamara's *Doing Well Enough* (DWE) framework developed in the late 1980s and in the 1990s, which is the first attempt to provide a model theoretic framework designed specially to account for supererogation and associated concepts of common sense morality, as well as being a sustained examination of these notions and the language used to express them. We also look at a later agent-evaluative expansion of the framework. In Section 6, we turn to other more recent work on supererogation, quickly sketching some of these developments. Finally, in Section 7, we take stock and briefly conclude the chapter.

Let me note that this handbook entry is extracted from a longer manuscript that became too unwieldy for such an entry, and I regret that this means that interesting work had to be passed over, including work of this author, and in other places natural developments and expansions of, as well as alternatives to, what is covered had to also be set aside. In various places I can just give a nod to other's work that we cannot cover here.

1 The Traditional Scheme & "Standard Deontic Logic"[1]

1.1 The Traditional Scheme[2]

The Traditional Definitional Scheme

The fundamental normative statuses of what I call "the *Traditional Scheme*" (TDS) are these five:

it is obligatory that (**O**B)	it is omissible that (**O**M)
it is permissible that (**P**E)	it is optional that (**O**P)
it is impermissible that (**I**M)	it is non-optional that (**N**O)[3]

The first three are familiar, but the fourth is widely ignored, the fifth has regularly been conflated with "it is a matter of *indifference* that ϕ" (more below), and the sixth, if mentioned at all, is derivatively conflated with non-indifference (or significance). Typically, one of the first two (but the third or fourth would work as well) is taken as basic, and the other five notions are defined in terms of it, but the last two cannot serve to define any of the first four. Following many expositions in modal and deontic logics, we'll take the necessity operator (*deontic* necessity here) as basic, and define the rest accordingly:

[1] The scare quotes indicate that "*Standard* Deontic Logic" and "SDL" function more as proper names than descriptions, but SDL has been extensively studied, and much work in deontic logic in cast in contrast to it.

[2] By the "Traditional Scheme", I am simply referring to a bit of unsystematic deontic folklore roughly exhausted by the mention of TDS plus DS and/or TTC below, along with a replacement rule. See [McNamara, 1990; McNamara, 1996a; McNamara, 1996b]. Below, I will suggest it amounts to the classical modal system ED.

[3] These abbreviations are non-standard mnemonics. We will be adding a number of other monadic operators to this set. "O" is routinely used instead of "**O**B", and often read as "It ought to be the case that", "P" in turn is often used instead of "**P**E", and "F" (for "forbidden") instead of "**I**M", and "I" is routinely used instead of "**O**P", and read as "It is a matter of indifference that". Deontic non-necessity, here denoted by "**O**M" is seldom ever named. The double lettering will also facilitate later discussion involving just what notions to take SDL and kindred systems to be modeling. Here we choose to read the basic operator as "it is obligatory that" so that continuity with permissibility, impermissibility, and optionality is not lost, as it would be with the "it ought to be the case that" reading. A choice must be made. "It is obligatory that" may also be read personally, but non-agentially as "it is obligatory for Jones that" [Krogh and Herrestad, 1996; McNamara, 2004])

(TDS)	$\mathbf{P}\mathrm{E}\phi \leftrightarrow \neg\mathbf{O}\mathrm{B}\neg\phi$	(Permissibility)
	$\mathbf{I}\mathrm{M}\phi \leftrightarrow \mathbf{O}\mathrm{B}\neg\phi$	(Impermissibility)
	$\mathbf{O}\mathrm{M}\phi \leftrightarrow \neg\mathbf{O}\mathrm{B}\phi$	(Omissiblity)
	$\mathbf{O}\mathrm{P}\phi \leftrightarrow (\neg\mathbf{O}\mathrm{B}\phi \;\&\; \neg\mathbf{O}\mathrm{B}\neg\phi)$	(Optionality)
	$\mathbf{N}\mathrm{O}\phi \leftrightarrow (\mathbf{O}\mathrm{B}\phi \vee \mathbf{O}\mathrm{B}\neg\phi)$.[4]	(Non-optionality)

Call this "*The Traditional Definitional* Scheme (TDS)". Although controversial, these equivalences are natural enough, and this scheme is still often employed, with the most focus on the first two definitions, and it is routinely presupposed in contexts of supererogation (although with the same conflation of *indifference* with *optionality* already mentioned, and to be discussed below).

The Traditional Threefold Classification, and the Deontic Square (and hexagon)

In addition to the TDS, it was traditionally assumed that the following, call it "The Traditional Threefold Classification" holds (Figure 1).

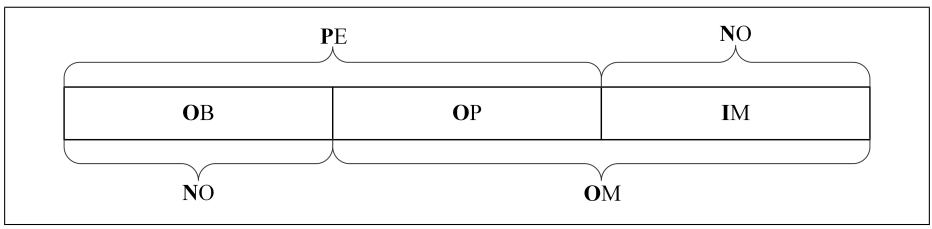


Figure 1: The Traditional Threefold Classification

The partition is the dark-lined figure. The three conditions indicated by the internal labels are intended to be mutually exclusive and jointly exhaustive: every proposition is either (overridingly) obligatory, optional, or impermissible, and no more than one of these. Let $\mathrm{MJ}^3(\mathbf{O}\mathrm{B}\phi, \mathbf{O}\mathrm{M}\phi, \mathbf{I}\mathrm{M}\phi)$ be shorthand for this formula:

(TTC) $(\mathbf{O}\mathrm{B}\phi \vee \mathbf{O}\mathrm{P}\phi \vee \mathbf{I}\mathrm{M}\phi) \;\&\; [\, \neg(\mathbf{O}\mathrm{B}\phi \;\&\; \mathbf{O}\mathrm{P}\phi) \;\&\; \neg(\mathbf{O}\mathrm{B}\phi \;\&\; \mathbf{I}\mathrm{M}\phi) \;\&\; \neg(\mathbf{O}\mathrm{P}\phi \;\&\; \mathbf{I}\mathrm{M}\phi)]$.[5]

[4] Here such equivalences will be called "definitions", sloughing over the distinction between definitional abbreviations and actual equivalence axioms encoding the force of such definitions.

[5] We will define MJ^n more carefully in Section 1.2. "MJ" is chosen as a mnemonic for "mutually exclusive and jointly exhaustive".

"The Deontic Square" (DS)" is also part of the Traditional Scheme (Figure 2).[6]

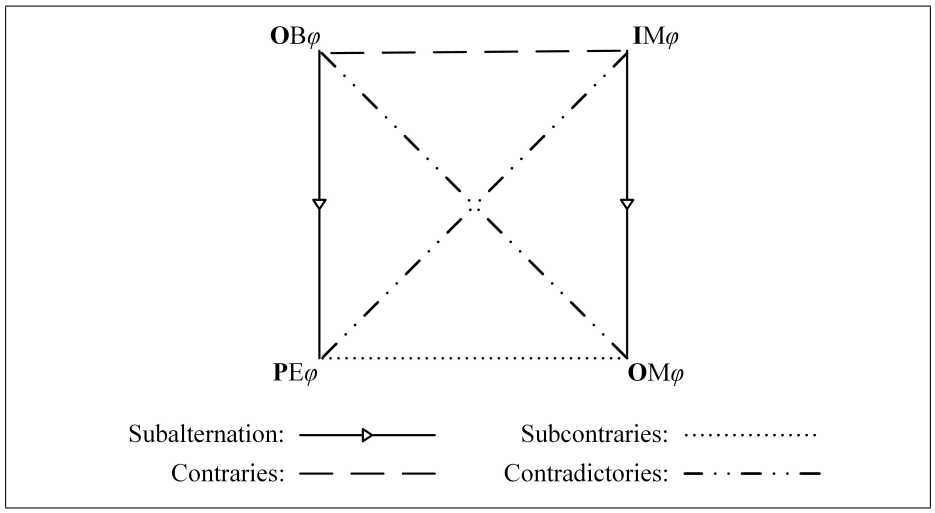


Figure 2: The Deontic Square

As a single formula, DS amounts to this:

(DS) $(\mathbf{O}\mathrm{B}\phi \leftrightarrow \neg\mathbf{O}\mathrm{M}\phi)$ & $(\mathbf{I}\mathrm{M}\phi \leftrightarrow \neg\mathbf{P}\mathrm{E}\phi)$ & $\neg(\mathbf{O}\mathrm{B}\phi$ & $\mathbf{I}\mathrm{M}\phi)$ & $(\mathbf{P}\mathrm{E}\phi \vee \mathbf{O}\mathrm{M}\phi)$ & $(\mathbf{O}\mathrm{B}\phi \rightarrow \mathbf{P}\mathrm{E}\phi)$ & $(\mathbf{I}\mathrm{M}\phi \rightarrow \mathbf{O}\mathrm{M}\phi)$

This square is perfectly analogous to one often displayed for the four alethic modalities, $\Box$, $\Box\neg$, $\Diamond$, and $\Diamond\neg$ (as well as classical quantifiers, among many others).[7] If we add nodes for OP (optionality) and ¬OP (non-optionality), we get a *deontic hexagon* (Figure 3).
An important logical feature of optionality is *the indifference of optionality to negation*:

(ION) $\mathbf{O}\mathrm{P}\phi \leftrightarrow \mathbf{O}\mathrm{P}\neg\phi$

ION follows from the TDS & RE (replacement of logical equivalents) assuming a classical propositional logic (PL): for $(\neg\mathbf{O}\mathrm{B}\phi$ & $\neg\mathbf{O}\mathrm{B}\neg\phi) \leftrightarrow$

[6]Recall the meaning of these oppositional relations: *contraries* cannot both be true, *subcontraries* cannot both be false, *contradictories* must have opposing truth values, and *subalternation* is the asymmetric relation of proper entailment–one item's entailing another, but not vice versa (e.g. the listing here of $\mathbf{O}\mathrm{B}\phi \rightarrow \mathbf{P}\mathrm{E}\phi$, is intended to convey that it is a logical truth that $\mathbf{O}\mathrm{B}\phi \rightarrow \mathbf{P}\mathrm{E}\phi$, but not so for the converse, $\mathbf{P}\mathrm{E}\phi \rightarrow \mathbf{O}\mathrm{B}\phi$).

[7]See [Moretti, 2009; Moretti, 2004].

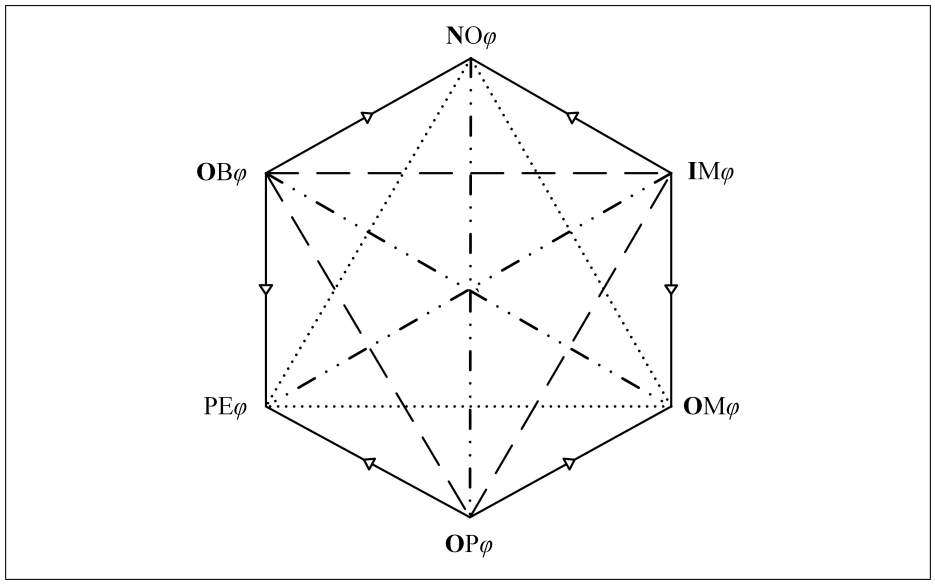


Figure 3: The Deontic Hexagon

$(\neg \mathbf{O}\mathrm{B}\neg\phi \;\&\; \neg \mathbf{O}\mathrm{B}\neg\neg\phi)$, so $\mathbf{O}\mathrm{P}\phi \leftrightarrow \mathbf{O}\mathrm{P}\neg\phi$.

1.2 From the Traditional Scheme to Standard Deontic Logic

The fundamental presupposition behind the Traditional Scheme

So what basic principles might the Traditional Scheme, with its TDS, TTC & DS, presuppose? First of all, the assumption is that the underlying logic is classical truth-functional propositional logic (henceforth, just "PL"). For example, $\mathbf{O}\mathrm{B}\phi \leftrightarrow \neg\mathbf{P}\mathrm{E}\neg\phi$ is endorsed, but this is definitionally equivalent to $\mathbf{O}\mathrm{B}\phi \rightarrow \neg\neg\mathbf{O}\mathrm{B}\neg\neg\phi$, which by PL is equivalent to $\mathbf{O}\mathrm{B}\phi \rightarrow \mathbf{O}\mathrm{B}\neg\neg\phi$, and the latter is *not* tautological. Here, clearly presupposed, is some principle of replacement of provable equivalents in the scope of *deontic* operators:

(RE) If $\phi \leftrightarrow \psi$ is a theorem, then so is $\mathbf{O}\mathrm{B}\phi \leftrightarrow \mathbf{O}\mathrm{B}\psi$

This is deemed one of the least controversial rules of inference for deontic logics, one characteristic of "classical modal logics" [Chellas, 1980]. With PL, RE and TDS, it is easy to prove the equivalences corresponding to the alternative definitional schemes mentioned above (e.g.

$\mathbf{O}\mathrm{B}\phi \leftrightarrow \neg\mathbf{P}\mathrm{E}\neg\phi$, taking now $\mathbf{P}\mathrm{E}$ as basic) and thus presupposed in the traditional scheme. What else?

The TTC and the DS expressed as formulae (above), given TDS, are easily shown to be each *tautologically* equivalent to the principle that (overriding) obligations cannot conflict,

(NC) $\neg(\mathbf{O}\mathrm{B}\phi \;\&\; \mathbf{O}\mathrm{B}\neg\phi)$[8]

and thus we record these meta-equivalences:

(EQ) Given TDS, formulas DS, TTC, and NC are tautologically equivalent to one another

Indeed, given TDS, TTC can be seen as a disguised version of NC along with RE:

(TD) The Traditional Scheme is essentially just the classical modal logic ED plus TDS[9]

As noted in [Hilpinen and McNamara, 2013, pp. 69-70], NC is not to be confused in content with

(OD) $\neg\mathbf{O}\mathrm{B}\bot$

OD asserts that no logical contradiction can be obligatory, whereas NC asserts that there can never be two things that are each separately obligatory, where the one obligatory thing is the negation of the other. The presence or absence of NC arguably represents one of the most fundamental divisions among deontic schemes, but it is a stronger claim than OD; and indeed almost all developments of logics for conflict-allowing obligations (i.e. those rejecting NC) accept OD.

As with much work in the history of normative ethics, early deontic logics presupposed that obligations could not conflict, or to put it more

[8] In primitive notation, DS is $(\mathbf{O}\mathrm{B}\phi \leftrightarrow \neg\neg\mathbf{O}\mathrm{B}\phi)$ & $(\mathbf{O}\mathrm{B}\neg\phi \leftrightarrow \neg\neg\mathbf{O}\mathrm{B}\neg\phi)$ & $\neg(\mathbf{O}\mathrm{B}\phi \;\&\; \mathbf{O}\mathrm{B}\neg\phi)$ & $\neg(\neg\neg\mathbf{O}\mathrm{B}\neg\phi \;\&\; \neg\neg\mathbf{O}\mathrm{B}\phi)$ & $(\mathbf{O}\mathrm{B}\phi \rightarrow \neg\mathbf{O}\mathrm{B}\neg\phi)$ & $(\mathbf{O}\mathrm{B}\neg\phi \rightarrow \neg\mathbf{O}\mathrm{B}\phi)$, and although the first two conjuncts are tautologies, the remaining four are each tautologically equivalent to NC above. Similarly, TTC becomes $(\mathbf{O}\mathrm{B}\phi \vee (\neg\mathbf{O}\mathrm{B}\phi \;\&\; \neg\mathbf{O}\mathrm{B}\neg\phi) \vee \mathbf{O}\mathrm{B}\neg\phi)$ & $[\neg(\mathbf{O}\mathrm{B}\phi \;\&\; \mathbf{O}\mathrm{B}\neg\phi)$ & $\neg(\mathbf{O}\mathrm{B}\phi \;\&\; (\neg\mathbf{O}\mathrm{B}\phi \;\&\; \neg\mathbf{O}\mathrm{B}\neg\phi))$ & $\neg((\neg\mathbf{O}\mathrm{B}\phi \;\&\; \neg\mathbf{O}\mathrm{B}\neg\phi) \;\&\; \mathbf{O}\mathrm{B}\neg\phi)]$, and the exhaustiveness clause is tautological, as are the last two conjuncts of the exclusiveness clause, but the first conjunct of that clause is just NC again. Likewise for the assumptions that $\mathbf{P}\mathrm{E}\phi \leftrightarrow (\mathbf{O}\mathrm{B}\phi \vee \mathbf{O}\mathrm{P}\phi)$ and $\mathbf{O}\mathrm{M}\phi \leftrightarrow (\mathbf{O}\mathrm{P}\phi \vee \mathbf{I}\mathrm{M}\phi)$.

[9] See [Chellas, 1980].

cautiously and plausibly, that the notion of obligations of importance in ethical theory did not allow for conflicts. Although it seems to this author, and many others, that it is obvious that there can be conflicting obligations, nonetheless, in the vast majority of work in ethics on supererogation, authors assumed they were dealing with a concept of obligation for which NC held. I think a safe course is to assume that when we read "**O**B" as "it is obligatory that" we should qualify this by adding an adjective that guarantees no conflicts: "overridingly". If it is overridingly obligatory that ϕ, then it is obligatory that ϕ and this obligation overrides all it conflicts with, and thus not only survives in the face of obligations it conflicts with, but it also defeats them, and thus renders them overridden. I submit that NC is *analytic* if we read the operator as "It is *overridingly* obligatory that". Let us do so unless otherwise stated henceforth. Theories allowing for conflicts and defeat and overriding of one obligation by another will have the resources to define this special subclass of obligations, and the derivability of NC should be a desideratum for success in expressing this notion of obligation.

Standard Deontic Logic (SDL)

Standard Deontic Logic can be seen as an expansion of the Traditional Scheme, motivated largely by analogies with (alethically interpreted) normal modal logics. Consider the following principle first:

(C) $(\mathbf{O}\mathrm{B}\phi \,\&\, \mathbf{O}\mathrm{B}\psi) \rightarrow \mathbf{O}\mathrm{B}(\phi \,\&\, \psi)$ (Aggregation of **O**B for &)

Early systems of deontic logic endorsed this principle, which says that if ϕ and ψ are each separately obligatory, then so too is their conjunction, ϕ & ψ. Although not entailed by TDS, we might say that it befits that scheme, since it is at least natural to think that if two things are each *overridingly* obligatory for me, then it is overridingly obligatory for me that both hold. The converse was also widely endorsed:

(M) $\mathbf{O}\mathrm{B}(\phi \,\&\, \psi) \rightarrow (\mathbf{O}\mathrm{B}\phi \,\&\, \mathbf{O}\mathrm{B}\psi)$ (Distribution of **O**B over &)

This does not seem to have a natural link to the Traditional Scheme, but it has a certain rationale behind it. M coupled with RE, is equivalent to this rule:

(RM) If $\phi \rightarrow \psi$ is a theorem,
then so is $\mathbf{O}\mathrm{B}\phi \rightarrow \mathbf{O}\mathrm{B}\psi$ (Inheritance)

RM allows us to make inferences from one thing's being obligatory to other things being obligatory, where those others are logical consequences of the former: if it is overridingly obligatory that I drive under 65 mph, then likewise for driving under 75 mph. This principle however is fully general (too general), and so also entails that logical truths are obligatory if anything is, since logical truths are entailed by anything. Thus in all but empty normative systems, $\mathbf{O}\mathrm{B}\top$ would hold. Since deontic logicians in the early years felt that empty normative systems could be set aside as uninteresting, they were ready to endorse not only that $\mathbf{O}\mathrm{B}\top$ would be true if anything was obligatory, but to treat it as a theorem for any logic of normative systems, thus endorsing it *simpliciter*:

(ON) $\mathbf{O}\mathrm{B}\top$

Together, we have the following rendering of SDL:

(A0) All propositional tautologies of the language
(A1) $\vdash \mathbf{O}\mathrm{B}\phi \rightarrow \neg\mathbf{O}\mathrm{B}\neg\phi$
(A2) $\vdash (\mathbf{O}\mathrm{B}\phi \ \& \ \mathbf{O}\mathrm{B}\psi) \rightarrow \mathbf{O}\mathrm{B}(\phi \ \& \ \psi)$
(A3) $\vdash (\mathbf{O}\mathrm{B}\phi \ \& \ \psi) \rightarrow (\mathbf{O}\mathrm{B}\phi \ \& \ \mathbf{O}\mathrm{B}\psi)$
(A4) $\vdash \mathbf{O}\mathrm{B}\top$
(MP) If $\vdash \phi \rightarrow \psi$ and $\vdash \phi$, then $\vdash \psi$
(RE) If $\vdash \phi \leftrightarrow \psi$, then $\vdash \mathbf{O}\mathrm{B}\phi \leftrightarrow \mathbf{O}\mathrm{B}\psi$

It is well known that this axiomatization of SDL is equipollent to the normal modal system D, which is often axiomatized as follows:

(A0) All propositional tautologies of the language
(A1′) $\mathbf{O}\mathrm{B}(\phi \rightarrow \psi) \rightarrow (\mathbf{O}\mathrm{B}\phi \rightarrow \mathbf{O}\mathrm{B}\psi)$
(A2′) $\mathbf{O}\mathrm{B}\phi \rightarrow \neg\mathbf{O}\mathrm{B}\neg\phi$
(MP) If $\vdash \phi \rightarrow \psi$ and $\vdash \phi$, then $\vdash \psi$
(RN) If $\vdash \phi$, then $\vdash \mathbf{O}\mathrm{B}\phi$

We will let "SDL" refer to either of the above systems, and for now, we will refer to "standard systems" as systems of deontic logic that contain SDL, perhaps expanding on SDL (e.g. by adding $\mathbf{O}\mathrm{B}(\mathbf{O}\mathrm{B}\phi \rightarrow \phi)$. We now turn briefly to the standard semantic treatment of SDL.

We give a standard "Kripke-style" possible world semantics for SDL. Assume that we have a set of possible worlds or situations, W, and a

binary relation, A, relating worlds to worlds. The intended reading is that Aij iff j is j is deontically *acceptable* from the standpoint of i, or more briefly, j is "i-acceptable" (so that no violations of the overriding obligations holding in i occur in j). We will denote i's acceptable worlds by A^i.[10] Formulae will be taken to be either true or false at a world, never both, and when a proposition, ϕ, is true at a world, we will often indicate this by referring to that world as a "ϕ-world". The truth-functional operators have their usual behavior at each world. Our focus will be on the contribution deontic operators are taken to make. The truth-condition for "**O**B" is as follows:

[OB] $\vDash_i \mathbf{O}\mathrm{B}\phi$ if and only if for every $j \in W$, if $\mathrm{A}ij$, then $\vDash_i \phi$

That is, it is obligatory that ϕ at i iff every i-acceptable world is a ϕ-world.
We add one constraint on the acceptability relation, namely that it is "serial": for every world, i in W, there is at least one world that is i-acceptable:

(SER) For every $i \in W$, there is a j in W such that Aij[11]

Following the treatment in normal modal logics, the fundamental idea here is that the deontic status of a proposition at a given world i is determined by how that proposition fairs at the i-acceptable worlds, as the diagrams in Figure 4 indicate. Here, we imagine that we gather together all the i-acceptable worlds, and then the status of a deontic formulae, $*\phi$ at i (where * is one of the six deontic operators) is determined by the status of ϕ in the i-acceptable worlds. The small dot represents the non-emptiness of A^i (SER). When a formula is true at every world in a model it is *true in the model*, and when a formula is true in every serial model, then the formula is *valid*. See Sections 6.1, A6.1 and A6.2 in [Hilpinen and McNamara, 2013] for a more full-bodied presentation

[10] The worlds related to i by A are also often called "ideal worlds", but we deliberately choose more neutral terminology, viewing the prevalent use of "ideal" as potentially misleading terminology, especially in the context of issues such as supererogation. (See [Hansson, 2006] for objections to invocation of ideal worlds in deontic logic in one of that term's senses.) Indeed, the choice of terminology in ethics and deontic logic has often tacitly contributed to the exclusion of supererogation from theorizing, and much confusion about it and other notions in the same family.

[11] Additional constraints on A will validate stronger logics than SDL itself. See Sections 7.1 and A7.1 of [Hilpinen and McNamara, 2013] for a sample and some references.

of SDL's syntax, proof theory, and semantics, and for further references.

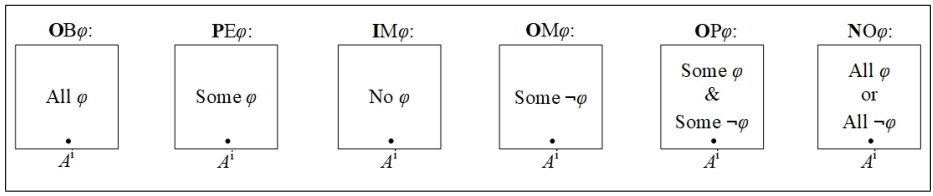


Figure 4: Truth Conditions for SDL Operators

2 Supererogation, indifference, and Urmson's constraint.

2.1 Moral Indifference, the Strong Threefold Classification, and Moral Rigor

Today I am wearing a pair of socks around the house, but often I go barefoot. To me, at this time of year, it is a matter of indifference. And as best I can tell, it is a matter of moral indifference as well. Now note that just as optionality is logically indifferent to negation, so too is moral indifference. For if it was a matter of moral indifference that I wear socks, then it is a matter of moral indifference whether I do or don't, and so it is also a matter of moral indifference that I don't wear socks. Thus we endorse "The Indifference of Indifference to Negation":

(IIN) $\mathbf{I}\mathrm{N}\phi$ iff $\mathbf{I}\mathrm{N}\neg\phi$

It should also be clear that the defining condition of optionality, namely, being neither obligatory nor impermissible, must be met by moral indifference. Thus we must endorse the "Optionality of Indifference":

(OI) $\mathbf{I}\mathrm{N}\phi \rightarrow \mathbf{O}\mathrm{P}\phi$

Note that *moral significance* can be plausibly defined via indifference (and vice versa):

(SI) $\mathbf{S}\mathrm{I}\phi \stackrel{\text{def}}{=} \neg\mathbf{I}\mathrm{N}\phi$

where "SI" is to be read as "it is a matter of significance that". Given the definition of significance, it is clear that NIS entails "The Indifference of

Significance to Negation":

(ISN) $\mathbf{SI}\phi \leftrightarrow \mathbf{SI}\neg\phi$

Reflection on the Deontic Hexagon also reveals that if we replace the two occurrences of "**O**P" and "**N**O" there with "**I**N" and "SI", all the resulting new logical links will also be intuitively sound (Figure 5).

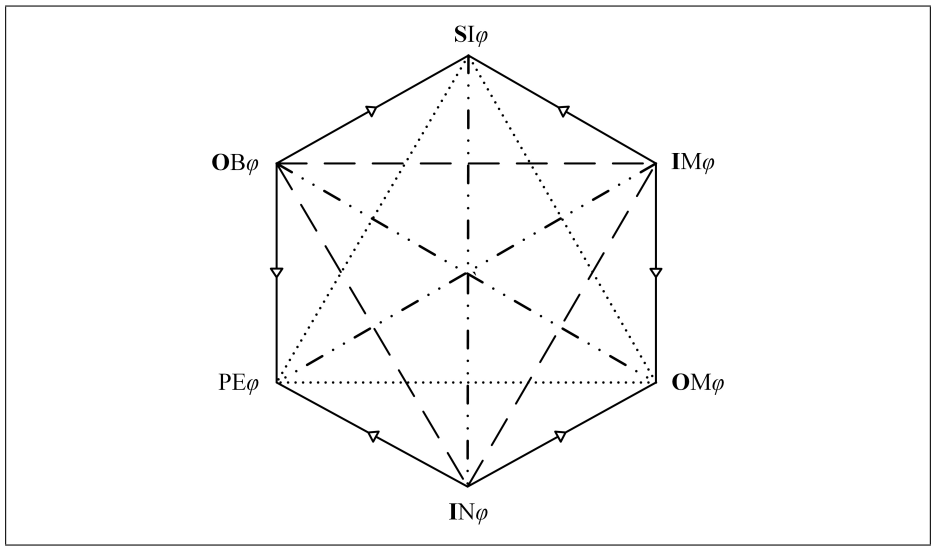


Figure 5: Alternate deontic hexagon

Now consider the following near-twin of TTC, call it the "*Strong* Threefold Classification:

(STC) $(\mathbf{O}\mathrm{B}\phi \vee \mathbf{I}\mathrm{N}\phi \vee \mathbf{I}\mathrm{M}\phi)$ & $[\neg(\mathbf{O}\mathrm{B}\phi$ & $\mathbf{I}\mathrm{N}\phi)$ & $\neg(\mathbf{O}\mathrm{B}\phi$ & $\mathbf{I}\mathrm{M}\phi)$ & $\neg(\mathbf{I}\mathrm{N}\phi$ & $\mathbf{I}\mathrm{M}\phi)]$

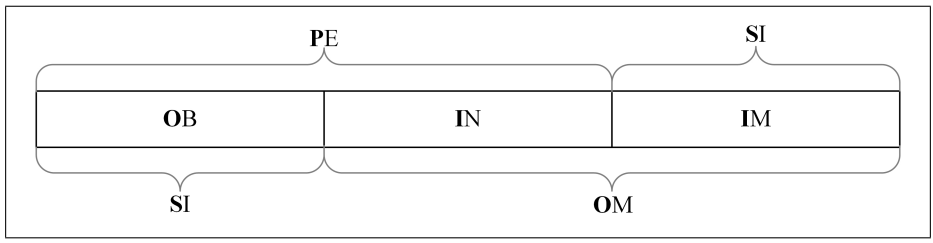


Figure 6: Strong Threefold Classification

Figure 6 provides a diagrammatic expression of STC.[12] Here indifference stands in for optionality, and as with the Traditional Threefold Classification, the claim is that each alternative action falls into one of the boxes, but no more than one. Note also that on the Traditional Scheme, STC is easily shown to be tautologically equivalent to what I will call "Moral Rigor":

(MR) $\quad \mathbf{O}\mathrm{P}\phi \leftrightarrow \mathbf{I}\mathrm{N}\phi$

Since we will need to explore a number of these mutually exclusive and jointly exhaustive claims for increasingly rich conceptual schemes, resulting in more complex formulas, let's make use of the following general shorthand henceforth:

$$\mathrm{MJ}^n(\mathrm{A}_1, \ldots, \mathrm{A}_n) \stackrel{\text{def}}{=} (\mathrm{A}_1 \vee \ldots \vee \mathrm{A}_n) \quad \& \; [(\mathrm{A}_1 \rightarrow \neg\mathrm{A}_2) \;\&\; (\mathrm{A}_1 \rightarrow \neg\mathrm{A}_3) \;\&\; \ldots \;\&\; (\mathrm{A}_1 \rightarrow \neg\mathrm{A}_n) \;\&\; (\mathrm{A}_2 \rightarrow \neg\mathrm{A}_3) \;\&\; (\mathrm{A}_2 \rightarrow \neg\mathrm{A}_4) \;\&\; \ldots \;\&\; (\mathrm{A}_2 \rightarrow \neg\mathrm{A}_n) \;\&\; \ldots \;\&\; (\mathrm{A}_{n-1} \rightarrow \neg\mathrm{A}_n)]^{13}$$

So $\mathrm{MJ}^1(\mathbf{O}\mathrm{B}\phi)$ is $\mathbf{O}\mathrm{B}\phi$, $\mathrm{MJ}^2(\mathbf{O}\mathrm{B}\phi,\mathbf{I}\mathrm{N}\phi)$ is $(\mathbf{O}\mathrm{B}\phi \vee \mathbf{I}\mathrm{N}\phi) \;\&\; (\mathbf{O}\mathrm{B}\phi \rightarrow \neg\mathbf{I}\mathrm{N}\phi)$, and $\mathrm{MJ}^3(\mathbf{O}\mathrm{B}\phi, \mathbf{I}\mathrm{N}\phi, \mathbf{I}\mathrm{M}\phi)$ is STC above (in feeble disguise).

2.2 Supererogation, indifference, optionality, and the Fivefold Classification[14]

Consider a dramatic case of supererogation.[15] An infant is trapped in a burning building. The fire has reached a very dangerous stage, sections of the building are in flames, windows are exploding, thick black smoke is pouring out of the entrance, etc. A mailwoman fond of the child, passes by and seeing a neighbor restraining the older sister, quickly sizes up the situation. Charging into the building and making her way to the top floor, she finds the infant still alive. On the verge of passing out,

[12] The outer labels of the diagram reflect these equivalences: $\mathbf{P}\mathrm{E}\phi \leftrightarrow (\mathbf{O}\mathrm{B}\phi \vee \mathbf{I}\mathrm{N}\phi)$ and $\mathbf{O}\mathrm{M}\phi \leftrightarrow (\mathbf{I}\mathrm{N}\phi \vee \mathbf{O}\mathrm{M}\phi)$ and $\mathbf{S}\mathrm{I}\phi \leftrightarrow (\mathbf{O}\mathrm{B}\phi \vee \mathbf{I}\mathrm{M}\phi)$.

[13] So "MJ"'s extension is a function, f, from numbers to truth-functions, and the extension of "MJ" followed by a numeral, "n", is that truth-function, $f(n)$, that maps n-tuples of truth values to true iff exactly one of the n values is true. (Thus the order of the truth values does not matter.)

[14] Later we will have cause to distinguish supererogation from action beyond the call of duty, but for now, we will follow the literature in not differentiating them.

[15] Cf. [Feldman, 1978, p. 46]

and badly burned, she lowers the child from a small shattered window and drops him to the neighbor below.

Our mailwoman's action is paradigmatic of the classical conception of supererogation. We can easily imagine the fire to be such that we would not even consider the firefighters to have been obligated to make such a direct-entrance rescue attempt. Yet we can also imagine that, although her action was very risky, it was not irresponsibly foolhardy [McNamara, 1996b]. Her action exceeded any demands morality made on her. She did more than she had to.

It is clear that the mailwoman's action was neither obligatory nor impermissible. Letting "SU" stand for "It is beyond the call (for Jane Doe) that" or "It is supererogatory that", these two features of supererogation can be jointly summed up as "The Optionality of Supererogation":

(OS) $\quad \mathbf{S}\mathrm{U}\phi \rightarrow \mathbf{O}\mathrm{P}\phi$

However, the classical conception of the supererogatory is obviously not exhausted by this feature. For despite the optionality of the mailwoman's action, her action was hardly a matter of moral indifference. Thus the classical conception supports "The Non-Indifference of Supererogation":

(NIS) $\quad \mathbf{S}\mathrm{U}\phi \rightarrow \neg\mathbf{I}\mathrm{N}\phi$

Together, the last two entailments yield "The Optional Non-Indifference of Supererogation":

(ONIS) $\quad \mathbf{S}\mathrm{U}\phi \rightarrow (\mathbf{O}\mathrm{P}\phi \ \& \ \mathbf{I}\mathrm{N}\phi)$

Let's also introduce in passing an operator that will be convenient to have later, for the non-supererogatory, the contradictory of supererogation:

(NS) $\quad \mathbf{N}\mathrm{S}\phi \stackrel{\mathrm{def}}{=} \neg\mathbf{S}\mathrm{U}\phi$

Recall that we saw earlier that we must endorse the "Optionality of Indifference", $\mathbf{I}\mathrm{N}\phi \rightarrow \mathbf{O}\mathrm{P}\phi$. But with supererogation in focus, it should now be apparent that the converse is problematic: to say that an action is indifferent is to say something stronger than that it is optional. We can easily imagine that it was a matter of moral indifference that our rescuer wore black socks that day or not, but not so for her rescu-

ing the infant, despite the fact that both were optional. So, replacing "**O**P" with "**I**N" in the Traditional Threefold Classification to yield the Strong Threefold Classification is quite contentious. STC simply makes too harsh a taskmaster of morality. For given ONIS, those committed to the possibility of supererogation, are thereby committed to what I'll call "Optionality with a Difference":

(OWD) $\mathbf{O}\mathrm{P}\phi \;\&\; \neg\mathbf{I}\mathrm{N}\phi$ is satisfiable

That is, it is possible that an alternative is both morally optional and morally significant. But STC simply rules out the possibility, and as noted above, it entails Moral Rigor, $\mathbf{O}\mathrm{P}\phi \leftrightarrow \mathbf{I}\mathrm{N}\phi$, thus enjoining the collapse of moral optionality and moral indifference. STC and MR each tacitly rule out the possibility of supererogation, and surely it is not the business of deontic logic to engage in such a substantive rejection of a pre-theoretic category.

It should also be clear that optionality cannot be equated with non-significance, since a supererogatory action is both morally optional and significant. Thus, although we must endorse the "Significance of the Non-Optional":

(SNO) $\mathbf{N}\mathrm{O}\phi \rightarrow \mathbf{S}\mathrm{I}\phi$

we must reject the converse, the "Non-Optionality of the Significant":

(NOS) $\mathbf{S}\mathrm{I}\phi \rightarrow \mathbf{N}\mathrm{O}\phi$

Let me plant here a question that may have already occurred to the reader: Given the semantic difference, yet logical overlap, between indifference and optionality (and significance and non-optionality), is there anything that distinguishes them at the level of logical principles? We will return to this later on, when we begin to look at semantic models for these notions.

So despite the fact that ethical theorists and deontic logicians have routinely, and often still do, label the condition of being neither obligatory nor impermissible as "indifference", this is an unwarranted and substantive conflation of two distinct and important deontic notions. As we will see, conflation of distinct pairs of concepts has been one major obstacle in finding a place for supererogation in deontic logic, as well as in ethical theory.

A preliminary extended classificational picture emerges as one to be expected (Figure 7).

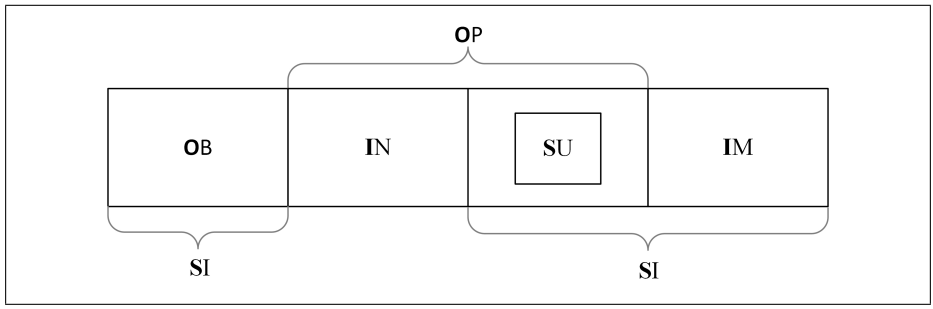


Figure 7: The Preliminary Fivefold Classification

Call it "The Preliminary Fivefold Classification", in symbols:

(PFC) $\quad \mathrm{MJ}^5(\mathbf{O}\mathrm{B}\phi, \mathbf{I}\mathrm{N}\phi, \mathbf{S}\mathrm{U}\phi\ [\mathbf{O}\mathrm{P}\phi\ \&\ \mathbf{S}\mathrm{I}\phi\ \&\ \mathbf{N}\mathrm{S}\phi], \mathbf{I}\mathrm{M}\phi)$

Reflection on supererogation and kindred notions forces this extension/enrichment of the TTC on us, whereas neglect of such notions encourages conflations of distinct concepts, as we have already seen in the case of indifference and optionality (and significance and non-optionality) with their conflation naturally leading to STC. These reflections also naturally lead to new questions that can generate new insights about expanded normative positions an agent might be in. For example, you may have wondered why the diagram above partitioned the non-indifferent optional alternatives into those that are supererogatory and those that are not. For if not, we could define the supererogatory actions as *the* non-indifferent optional actions, by adding to OIS, an endorsement of its converse "The Sufficiency of Optional Non-Indifference for Supererogation"

(SONS) $\quad (\mathbf{O}\mathrm{P}\phi\ \&\ \neg\mathbf{I}\mathrm{N}\phi) \rightarrow \mathbf{S}\mathrm{U}\phi$

We are now in a good position to see why this is unacceptable by the logical features of the concepts alone. Suppose $\mathbf{S}\mathrm{U}\phi$. By ONIS, $\mathbf{O}\mathrm{P}\phi$ & $\neg\mathbf{I}\mathrm{N}\phi$. Then by ION and IIN, $\mathbf{O}\mathrm{P}\neg\phi$ & $\neg\mathbf{I}\mathrm{N}\neg\phi$, and then by SONS, we get $\mathbf{S}\mathrm{U}\neg\phi$, thereby generating "The Indifference of Supererogation to Negation":

(ISUN) $\quad \mathbf{S}\mathrm{U}\phi \rightarrow \mathbf{S}\mathrm{U}\neg\phi$

ISUN is surely absurd, for it says the negation of what is supererogatory is supererogatory, and it entails that for any alternative, either it is not supererogatory or *both* it and its negation are.[16] Indeed, it is plausible to think that something stronger holds: that there can be "No Supererogatory Conflicts":

(NSC) $\neg(\mathbf{S}\mathrm{U}\phi \;\&\; \mathbf{S}\mathrm{U}\neg\phi)$

For suppose that ϕ is supererogatory for Jane Doe, and that ϕ consists of her doing A. Then her doing A while doing only permissible things must guarantee doing more than the minimum. But then presumably there must be permissible ways of her not performing A that don't involve doing more than the minimum. So even assuming that she does nothing but permissible things, A's non-performance can't assure that she has done anything beyond the minimum, and hence $\neg\phi$, her not doing A, can't be supererogatory. For example, surely our mailwoman's *not* rescuing the infant would not be supererogatory per se, for she might accomplish that by fanning the flames, or by just walking by without interfering with the situation in any way, or by merely helping direct the fire truck to the seen from the corner (the minimum, we imagine). So we do need to leave space for optional non-indifferent alternatives that are *not* supererogatory.

I now turn to a scheme that has been routinely confused with the traditional scheme, by friend and foe of supererogation, and has been a source of much confusion, as well as mistaken criticisms of deontic logic.

2.3 Urmson, friends of supererogation, and the Traditional Scheme

Most ethicists and deontic logicians have routinely and unreflectively endorsed "Moral Rigor",

(MR) $\mathbf{O}\mathrm{P}\phi \leftrightarrow \mathbf{I}\mathrm{N}\phi$

by taking "the morally indifferent" as the deontic analogue of contingency: as anything that is neither obligatory nor whose negation is obligatory. As we've seen, this conflates indifference with optionality. And this mistake leads to other conflations. For once the conflation of

[16] Indeed, as we shall see, a number of analysts have thought that if ϕ is supererogatory, then its absence, $\neg\phi$, is an offence or suberogatory, and so has a *negative* valence. Although we will reject this too, it does underscore the counter-intuitiveness of SONS.

optionality with indifference occurs it is but a short step on the traditional framework to conflating the Traditional Threefold Classification with its near twin, the "Strong Threefold Classification",

(STC) $\mathrm{MJ}^3(\mathbf{O}\mathrm{B}\phi, \mathbf{I}\mathrm{N}\phi, \mathbf{I}\mathrm{M}\phi)$

Although the Traditional Threefold Classification is relatively innocuous (when **O**B is read as "it is overridingly obligatory that"), STC is anything but. For consider this crucial component of STC, "Strong Exhaustion":

(SE) $\mathbf{O}\mathrm{B}\phi \vee \mathbf{I}\mathrm{N}\phi \vee \mathbf{I}\mathrm{M}\phi$

SE entails that if an alternative is neither obligatory nor permissible, then it must be a matter of moral indifference.[17] This is not to be confused with the relatively innocuous component of TTC, call it "Weak Exhaustion":

(WE) $\mathbf{O}\mathrm{B}\phi \vee \mathbf{O}\mathrm{P}\phi \vee \mathbf{I}\mathrm{M}\phi$

For SE, unlike WE, entails that morality rules with an "iron fist": for any alternative that is not a matter of moral indifference, morality will either demand that it hold or demand that it not hold. There simply are no morally significant optional alternatives according to SE.

Conflation of SE with WE, and optionality with indifference, has led to another recurring mistake. "The threefold scheme must go!" has often been the battle cry of the friends of supererogation in the polemical literature in ethical theory on supererogation ever since Urmson's classic [Urmson, J.O., 1958]. Now clearly, SE, and so STC, do rule out any possibility of supererogation by ONIS. So ONIS is rightly used by the friends of supererogation to place the onus on those who support STC. Notice however that as an argument against WE or the *Traditional* Threefold Classification, this is just a non sequitur. Despite all claims to the contrary, moral indifference is not even representable in the Traditional Scheme. So when friends of supererogation rally behind "the threefold scheme must go", they should be referring exclusively to SE or

[17] The other component, "Strong Mutual Exclusiveness", that no alternative falls into more than one of these categories, the morally obligatory, the morally impermissible or the morally indifferent, is plausible, especially given that we are reading "morally obligatory" as short for "*overridingly* morally obligatory". Similarly, for the exclusiveness component of TTC.

to the *Strong* Threefold Classification. However, the distinction between the strikingly similar three-fold classifications is rarely made. Ethicists and deontic logicians alike are routinely guilty of conflating TTC with the STC as a result of conflating moral indifference with moral optionality.[18] Then the friends of supererogation, inherit and propagate these mistakes in the process of trying to fight for a place for supererogation. In deontic logic, this confusion goes right back to the beginning of deontic logic as an active ongoing area of research, [von Wright, 1951, p. 3]. Ironically, even in Urmson's own classic on supererogation, we find him conflating indifference with optionality–even as he himself was leading the way in our achieving escape velocity from the conflation. But his intention was clear.[19] Any scheme that entails the *Strong* Threefold Classification (or Moral Rigor) is inconsistent with the possibility of supererogation. So we can take "Urmson's (general) Constraint" on deontic schemes to be:

(UC) $\mathbf{I}\mathrm{N}\phi \rightarrow \mathbf{O}\mathrm{P}\phi$ is a logical truth, but $\mathbf{O}\mathrm{P}\phi \rightarrow \mathbf{I}\mathrm{N}\phi$ is not.

3 The early axiology-based work of Chisholm

In a series of papers in the mid-sixties, Chisholm, and Chisholm-and-Sosa, provide conceptual schemes, using axiological notions as foundational, that aimed to make a place for supererogation and kindred notions. The most famous and influential of these by far is "Supererogation and Offence: A Conceptual Scheme for Ethics" [Chisholm, 1963b]. My main focus in this section will be on this piece, but I will briefly consider in passing the work of Chisholm-Sosa [Chisholm and Sosa, 1966b; Chisholm and Sosa, 1966a], but I must reserve non-cursory coverage of that joint work for another place.[20]

[18] This charge is defended explicitly in [McNamara, 1990], and also in unpublished presentations [McNamara, 1988; McNamara, 2006]). Chisholm, whose work we will examine shortly, is an exception.

[19] Urmson's conflation is also discussed explicitly in [McNamara, 1990], and in unpublished presentations [McNamara, 1988; McNamara, 2006]). Even Chisholm himself makes the same mistake in his seminal "Supererogation and Offence" [Chisholm, 1963b, pp. 326-27], which contains all the ingredients needed to see the difference and recognize that the traditional deontic scheme of deontic logic is none other than TTC, even if the explicit defining condition, $\neg\mathbf{O}\mathrm{B}\phi$ & $\neg\mathbf{I}\mathrm{M}\phi$, is mislabeled as "indifferent".

[20] The Chisholm-Sosa work is closer to [Chisholm, 1963b] in various ways, including taking an axiological stance as basic. In the next section we will look at [Chisholm, 1964], [Chisholm, 1974] which develops Chisholm's thinking in a different direction, this time about prima facie obligation, defeat, etc.

3.1 Chisholm's "Supererogation and Offence"

[Chisholm, 1963b] is the most important short piece on supererogation since Urrmson's classic [Urmson, 1958]. It is rich in ideas and insights, as well as being informative about the related work of Meinong. It can be helpful to see him in retrospect, as in part, attempting to identify and argue for an enriched set of *normative position*, as well as providing an analytic framework for these normative positions.[21] Chisholm's piece is also quite important for its introduction and discussion of a negative analog of supererogation (now often called "suberogation") in Anglo-American philosophy, a proposed analog of continuing controversy. Finally, he makes a variety of preliminary and critical points about views held by others or views that he thought might naturally arise about the nature of supererogation and kindred notions, and many of these critical points have been widely endorsed.

Supererogation

Preliminary to providing his proposed analytic framework, he asks us to "take the supererogatory to be that which it is good but not obligatory to do, in short, "non-obligatory well-doing" [Chisholm, 1963b, p. 3][22]. Fol-

[21] In deontic logic, this typically involves an exhaustive account relative to a deontic logic of a person's normative relationship to a proposition or a person's normative relationship to an agential proposition–a proposition that attributes agency to a person regarding another proposition (the person may even be left implicit when exploring singular positions.) See [Sergot, 2013] in the 1st volume of this handbook for a historical and systematic overview by the leading figure in this area of research. Chisholm's discussion is cast freely in terms of actions and an agent's relationship to actions more than propositions however. We will recast things here for simplicity.

[22] We will see that these characterizations are fundamentally inadequate, as are other similar subsequent analyses. However, let me note here a minor clarification: Chisholm himself will later speak of what *would be* good and *would be* bad when introducing his own scheme, so we should read his glosses that way in general, so that "it is good that" should either be read as "it would be good that" or "it is a good *option* that" in discussing Chisholm. This is important, for as [Goble, 1990a] rightly notes, on its face, "it is good that p" is factive and thus entails p. Similar remarks apply to "it is bad that". Goble develops an actualist deontic logic for *good, bad, better*, and *ought* in a series of impressive papers: [Goble, 1990a; Goble, 1990b; Goble, 1989]. (Note that the 1989 paper builds on the two 1990 papers and was written after those.) The behavior of the operators for good and bad in the context of truth functional connectives is fully articulated for the actualist semantic framework Goble articulates. However, supererogation and such is certainly not in focus in those papers, and in particular, I do not think that the possibility of non-equally good but mutually exclusive good or bad options fits well, nor for example, good but

lowing Urmson, Chisholm notes that some authors have assumed that all of an agent's alternatives must fall into one of three mutually exclusive and exhaustive classes, the obligatory, the forbidden and the *indifferent* (see STC above)[23], but that supererogatory acts don' fit into any of these three classes. Following both Urmson and Ladd, he notes that just as there can be trifling obligations (e.g. returning a pen), and highly meritorious duties involving the sacrifice of one's life (e.g. for a person holding a vital position facing an enemy onslaught), there can likewise be trifling and highly meritorious supererogatory acts (e.g. contrast a small favor, with a mailwoman's risking her life to save some child.[24] He then draws some conclusions about supererogatory acts ([Chisholm, 1963b], p.3-5): (i) supererogatory acts are not necessarily better than or more morally praiseworthy than acts of duty; (ii) the performance of a supererogatory act needn't imply or reflect any standing virtue of the agent (e.g. a selfish person may have a charitable out-of-character moment); (iii) nor "can the difference between duty and supererogation be made out by reference to the traditional distinction between those duties which are 'perfect' and those which are 'imperfect'.[25]

Offence (suberogation)

Chisholm goes on to famously (or some might say, infamously)[26] argue that supererogation has a natural analog forced on us by parity of reason,

non-obligatory options, or bad to not do but not obligatory options (e.g. offensive omissions), since the framework validates thesis like $GD\phi \leftrightarrow (OU\phi \ \& \ \phi)$, and $OU\phi \leftrightarrow (GD\phi \vee BD\neg\phi)$. However, it is also not clear this matters, since Goble explicitly interprets "O" in these papers as *it ought to be that*, and so is not intending to link the goodness/badness of what an agent brings about to what an agent is obligated to do, which does seem to be Chisholm's main focus. However, this does raise a general question about the relationship of *ought* to supererogation: if, as is very plausible, it can be good to do something less than the best one could do, then can it be that if done, it ought to be that it was done given that it rules out the best option available? The general relationship between *ought* and supererogation, which can be and has been perplexing, will be explored more later in the chapter.

[23]Chisholm cites logicians work in deontic logic: [von Wright, 1951; von Wright, 1953], [Prior, 1962 1955](the 1955 edition), and [Anderson, 1956].

[24]The examples do not all match those in Chisholm's text, but are in keeping with the intent.

[25]Although I think the argument he gives against this is defective (more below), I believe the point is sound.

[26]Even the staunchest defenders of supererogation (see [Heyd, 1982], [Mellema, 1991]) raise serious doubts about offences and argue against the alleged symmetry between offence and supererogation.

"what is bad but not forbidden", or "permissive ill-doing" [Chisholm, 1963b, p. 5]. He tells us that

> "a system of moral concepts which provides a place for what is good but not obligatory [supererogatory], should also provide a place for what is bad but not forbidden. For if there is such a thing as 'non-obligatory well-doing' then it is plausible to suppose that there is also such a thing as 'permissive ill-doing"'.

Chisholm appropriates the term "offence" for this sort of act.[27] He offers two sorts of examples of offences as negative analogs to the prior cases of trifling supererogation and of highly meritorious supererogation: a) a trifling offence consisting of an act of discourtesy and b) a villainous offence consisting of the sort of acts he thinks informers often engage in, for example someone who permissibly provides very damaging information about a competitor for self-interested and malicious reasons.[28] Thus was introduced into Anglo-American philosophy, a new category in ethical theory, as well as a claim that has been much disputed ever since: that there is a perfectly symmetric negative analog to supererogation.[29] Chisholm goes on to note that, just as with acts of supererogation, offences do not fit into the then-widely endorsed STC, for permissible bad actions are also neither obligatory, nor forbidden, nor indifferent. Chisholm then turns to a brief discussion of an earlier improvement on this STC scheme, one not yet then noted in Anglo-American philosophy or deontic logic.

[27]The British spelling 'offence', rather than the American 'offense', has predominated since Chisholm used it here, and I follow that tradition, while also using 'offence' and 'suberogation' interchangeably (but see [McNamara, 2011b] for more nuance on "offence").

[28]These examples are hardly uncontroversial, but we must let this matter pass here. See [Driver, 1992] for a defense of the viability of the category.

[29]See for example discussion in [Heyd, 1982], [Mellema, 1991], [Driver, 1992], [Zimmerman, 1996], and [Heyd, 2012]). One aspect of the symmetry of traditional conceptions of supererogation and suberogation is discussed in [McNamara, 2011b],

Chisholm on Meinong's Scheme[30]

Chisholm attributes the following five fold classificational scheme to Alexius Meinong[31]: Every act is either indifferent, meritorious, required, excusable, or reprehensible, but no more than one of these. Taking the permissible to be what is not reprehensible, and the optional to be that which is neither required nor reprehensible, the following picture emerges (Figure 8).

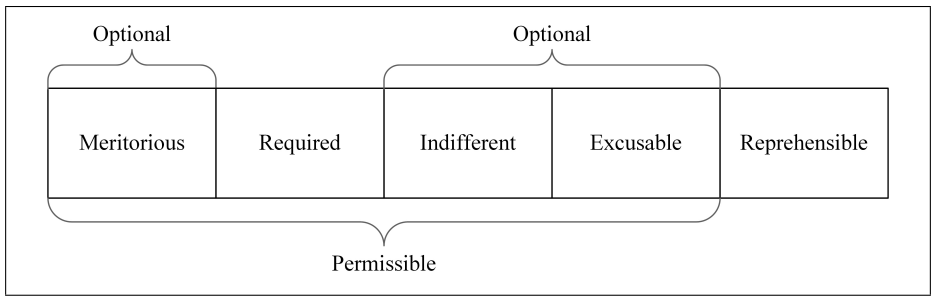


Figure 8: Meinong's five-fold classification

Chisholm suggests that given Meinong's examples and claims about his categories, he can map Meinong's terminology on to his as follows:

Meritorious – Supererogatory	Required – Obligatory
Excusable – Offence	Reprehensible – Forbidden

Letting "**O**F" stand for "it is an offence that", and letting our "**I**M" (for "impermissible") stand in for Chisholm's "Forbidden", we get the partition in Figure 9.

[30]Because of the singular influence of Chisholm's article, my focus here is on Chisholm's own account of Meinong and its influence on his scheme. Although I have no reason to think there are egregious errors in Chisholm's account, I would like to direct the reader's attention to Chapter one, Section 2 of the first volume of this handbook [Hilpinen and McNamara, 2013]. It certainly overlaps in content with Chisholm's discussion of Meinong, but also contains more, and of course other references to work on Meinong and deontic logic that came after Chisholm, as well as references to older Islamic work with affinities to Meinong's scheme; and let me add [Purtill, 1973].

[31]Chisholm cites [Meinong, 1894] primarily, but he also cites [Meinong, 1968] as providing further details on laws of omission (more on this below). See also [Hilpinen and McNamara, 2013, Section 2, 9-15] in the first volume of this handbook, and references there.

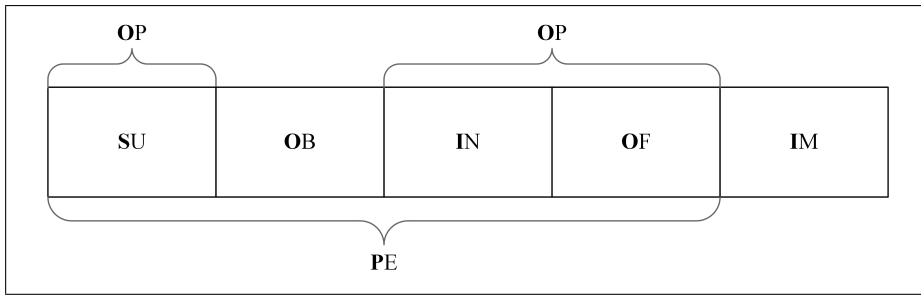


Figure 9: Meinong's Fivefold Classification via operators

Cast symbolically via propositional operators, Meinong's Fivefold Classification is:

(MFC) $\quad \mathrm{MJ}^5(\mathbf{S}\mathrm{U}\phi, \mathbf{O}\mathrm{B}\phi, \mathbf{I}\mathrm{N}\phi, \mathbf{O}\mathrm{F}\phi, \mathbf{I}\mathrm{M}\phi)$

Letting "**G**D" stand for "it is good that", and "**B**D" for "it is bad that", Chisholm tells us that Meinong also endorsed what I will call Meinong's *Deontic-Axiological* bridge principles:

(MD-A) $\quad (\mathbf{S}\mathrm{U}\phi \vee \mathbf{O}\mathrm{B}\phi) \rightarrow \mathbf{G}\mathrm{D}\phi$
$\quad\quad\quad\quad (\mathbf{O}\mathrm{F}\phi \vee \mathbf{I}\mathrm{M}\phi) \rightarrow \mathbf{B}\mathrm{D}\phi$.[32]

Adding to the categories diagram, we get Figure 10.

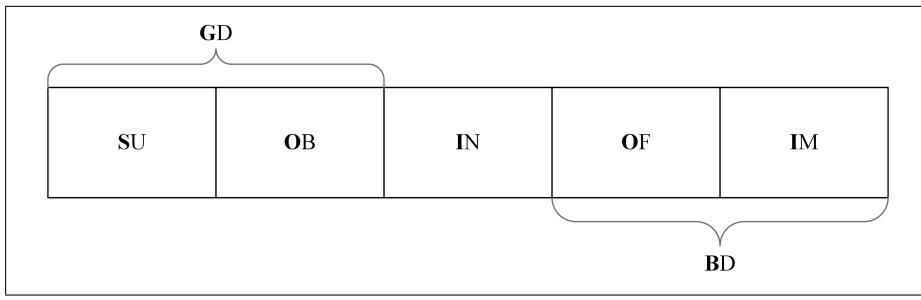


Figure 10: Meinong's five-fold classification and *good* and *bad*

Secondly, Chisholm tells us that Meinong also endorsed a thesis to the effect that the categories above are ranked left to right in descending order of value. Letting "$>$" stand for a ranking relation, "Meinong's Ranking Thesis" is:

[32]Chisholm does not say whether the converses are intended by Meinong. We will see that these are retained in Chisholm's own scheme.

(MRT) $(\mathbf{S}\mathrm{U}\phi \;\&\; \mathbf{O}\mathrm{B}\psi \;\&\; \mathbf{I}\mathrm{N}\chi \;\&\; \mathbf{O}\mathrm{F}\lambda \;\&\; \mathbf{I}\mathrm{M}\theta) \rightarrow$
$(\phi > \psi > \chi > \lambda > \theta)$

Finally, he tells us that Meinong endorsed the following *Laws of Omission* (where A stands for Jane Does performing some action, A):

(MLO) a) $\mathbf{S}\mathrm{U}A \leftrightarrow \mathbf{O}\mathrm{F}\neg A$
b) $\mathbf{S}\mathrm{U}\neg A \leftrightarrow \mathbf{O}\mathrm{F}A$
c) $\mathbf{O}\mathrm{B}A \leftrightarrow \mathbf{I}\mathrm{M}\neg A$
d) $\mathbf{O}\mathrm{B}\neg A \leftrightarrow \mathbf{I}\mathrm{M}A$

Chisholm objects to MRT and to MLO. Against MRT, he notes that his earlier points apply, that a highly meritorious obligatory act can be better than a trifling supererogatory favor, and that a trifling impermissible act might be better than a heinous offence. He accepts the last two laws of omission but rejects the first two by elaborating on one of Meinong's own examples to illustrate their implausibility. Even if winning by near-cheating is an offence, it doesn't follow that the bare omission of winning by near-cheating is supererogatory. Similarly, if intentionally setting aside a permissible gain on behalf of a competitor is supererogatory, it doesn't follow that any bare omission of doing so is an offence.

Chisholm's own conceptual scheme

Chisholm notes that MLO provides an important insight, one to be carried forward in his own scheme: *in considering how to classify the status of a given action, we must consider both the status of its performance and the status of its non- performance.* Meinong also provides a possible clue for classifying the statuses of actions that Chisholm adopts: *employ two contrary terms in combination to define the target normative concepts.* Chisholm lists a number of such pairs and selects *good*, and *bad*, asking us to interpret "good" as in "that would be a good thing to do", and "bad" as in "that would be a bad thing to do", stressing that it is the thing done, not the agent or consequences that the pair applies to [Chisholm, 1963b, p. 10].[33] However, it is clear that what he has in

[33]Chisholm does not say why he chooses the last pair. Notice that there is an ambiguity in "non-performance" that is revealed by asking "Is anything that is not the performance of an action, a non-performance of any given action? If so, then the sun's rising tomorrow is a non-performance. This is probably not what Chisholm meant, but on the other hand, the final section of the article strongly favors the idea that "good" and "bad" apply to states of affairs or propositions generally, and need

mind is both performances *and non-performances* of actions.
For Chisholm next tells us that for each action, there are three possibilities:

1) it would be good,
2) it would be bad,
3) it would be neither good nor bad,

and he immediately goes on to note that (1) - (3) above apply to both performances *and non-performances* of actions. Finally, he notes that we must take "good" and "bad" so that "their application to performance is logically independent of their application to non-performance" [Chisholm, 1963b, p. 10]. This is ambiguous, but he indicates explicitly that he means that *being good to do* does not entail *being bad to not do*, and vice versa. We will return to this ambiguity.

With these preliminaries, he then notes that we get the following nine logically possible combinations of *good* and *bad* for the *performance* and corresponding *non-performance* of any action, and he proposes a new conceptual scheme in terms of these via the labels on the right:

	performance value:	non-performance value:	type of act:
1.	good	good	Totally Supererogatory
2.	good	neither	Supererogatory Commission
3.	good	bad	Obligatory
4.	neither	good	Supererogatory Omission
5.	neither	neither	Indifferent
6.	neither	bad	Offence of Omission
7.	bad	good	Impermissible
8.	bad	neither	Offence of Commission
9.	bad	bad	Totally Offensive

3.2 Regimentation of Chisholm's scheme and assessment

Syntax for RCGB logics

Expressive resources & Chisholm's Definitional Scheme (CDS)
In this section, we will introduce a preliminary reconstruction of what a logic for Chisholm's scheme might look like. We will work our way to this by a bit of reverse engineering.

not involve an agent, and as mentioned above, the joint work with Sosa clearly does this, as does Chisholm's work on the logic of requirement. (See Section 4.)

Chisholm does not say anything about how the two primitives might apply to compounds of performances and/or non-performances of actions by agents. I proposed that we explore a regimentation of Chisholm's approach, here casting it in propositional form,[34] with an interpretation of the atomic sentences that perhaps best matches, in a propositional context, what Chisholm had in mind. Assume that we have a restricted language, where say any atomic sentence, P_i, attributes an action to our mock agent, Jane Doe, so that $\neg\mathrm{P}_i$ would then express the claim that it is not the case that Jane Doe did perform that action, and that we have **G**D and **B**D again as propositional operators.[35] With this understanding in mind, we define the formulae.

RCGB formulas:

1) $P_1, \ldots, P_n, \ldots$ are RCGB formulas;
2) $\bot$ and $\top$ are RCGB formulas;
3) If ϕ is an RCGB formula, so is $\neg\phi$, $\mathbf{G}\mathrm{D}\phi$, and $\mathbf{B}\mathrm{D}\phi$;
4) If ϕ and ψ are RCGB formulas, so are $(\phi \vee \psi)$, $(\phi \,\&\, \psi)$, $(\phi \rightarrow \psi)$, and $(\phi \leftrightarrow \psi)$.

We recast Chisholm's Definitional Scheme as follows:

[34] In [Chisholm and Sosa, 1966a], [Chisholm and Sosa, 1966b] they treat **G**D and **B**D as operators, or at least close surrogates thereof (properties of states of affairs), and not restricted to actions or non-actions.

[35] We could read the operators as qualifying how Jane Doe *is to be* (e.g. It is Good/Bad/Neutral/Obligatory/ Supererogatory ... *for Jane Doe to be such that* ϕ. This is a personal, but non-agential reading. ([McNamara, 2004] argues that in the case of obligation at least, this construction can be used, along with an agency operator to define agential obligations as a special case. It would be easy enough to add in an agential operator, and go on to map out the normative agential positions. However, since Chisholm speaks of performance of actions and nonperformance of actions, it would also be interesting to explore what his scheme might look like if formally recast in a language that represented actions, and their performance and non-performance. See Section 9 of Chapter 1 of the first volume of this handbook [Hilpinen and McNamara, 2013] for a brief survey of some approaches to the logic of action and agency. Here we keep matters simple and slough over some subtleties. Lastly, we might also cast this in the language of modal agency via a "brings it about that" operator.

(CDS)	
Totally Supererogatory:	$\mathbf{TS}\phi \stackrel{\text{def}}{=} \mathbf{G}\mathrm{D}\phi \;\&\; \mathbf{G}\mathrm{D}\neg\phi$
Supererogatory Commission:	$\mathbf{SU}\phi \stackrel{\text{def}}{=} \mathbf{G}\mathrm{D}\phi \;\&\; \neg(\mathbf{G}\mathrm{D}\neg\phi \vee \mathbf{B}\mathrm{D}\neg\phi)$
Obligatory:	$\mathbf{OB}\phi \stackrel{\text{def}}{=} \mathbf{G}\mathrm{D}\phi \;\&\; \mathbf{B}\mathrm{D}\neg\phi$
Supererogatory Omission:	$\mathbf{SU}\neg\phi \stackrel{\text{def}}{=} \mathbf{G}\mathrm{D}\neg\phi \;\&\; \neg(\mathbf{G}\mathrm{D}\phi \vee \mathbf{B}\mathrm{D}\phi)$
Indifferent:	$\mathbf{IN}\phi \stackrel{\text{def}}{=} \neg(\mathbf{G}\mathrm{D}\phi \vee \mathbf{B}\mathrm{D}\phi) \;\&\; \neg(\mathbf{G}\mathrm{D}\neg\phi \vee \mathbf{B}\mathrm{D}\neg\phi)$
Offence of Omission:	$\mathbf{OF}\neg\phi \stackrel{\text{def}}{=} \mathbf{B}\mathrm{D}\neg\phi \;\&\; \neg(\mathbf{G}\mathrm{D}\phi \vee \mathbf{B}\mathrm{D}\phi)$
Impermissible:	$\mathbf{IM}\phi \stackrel{\text{def}}{=} \mathbf{B}\mathrm{D}\phi \;\&\; \mathbf{G}\mathrm{D}\neg\phi$
Offence of Commission:	$\mathbf{OF}\phi \stackrel{\text{def}}{=} \mathbf{B}\mathrm{D}\phi \;\&\; \neg(\mathbf{G}\mathrm{D}\neg\phi \vee \mathbf{B}\mathrm{D}\neg\phi)$
Totally Offensive:	$\mathbf{TO}\phi \stackrel{\text{def}}{=} \mathbf{B}\mathrm{D}\phi \;\&\; \mathbf{B}\mathrm{D}\neg\phi$

If we allow the following shorthand, "**N**U" for "it is neutral that"

$$\mathbf{N}\mathrm{U}\phi \stackrel{\text{def}}{=} \neg(\mathbf{G}\mathrm{D}\phi \vee \mathbf{B}\mathrm{D}\phi),$$

the categorical scheme can be expressed more concisely:

(CDS)	status of: ϕ: $\neg\phi$:
Totally Supererogatory:	$\mathbf{TS}\phi \stackrel{\text{def}}{=} \mathbf{G}\mathrm{D}\phi \;\&\; \mathbf{G}\mathrm{D}\neg\phi$
Supererogatory Commission:	$\mathbf{SU}\phi \stackrel{\text{def}}{=} \mathbf{G}\mathrm{D}\phi \;\&\; \mathbf{N}\mathrm{U}\neg\phi$
Obligatory:	$\mathbf{OB}\phi \stackrel{\text{def}}{=} \mathbf{G}\mathrm{D}\phi \;\&\; \mathbf{B}\mathrm{D}\neg\phi$
Supererogatory Omission:	$\mathbf{SU}\neg\phi \stackrel{\text{def}}{=} \mathbf{N}\mathrm{U}\phi \;\&\; \mathbf{G}\mathrm{D}\neg\phi$
Indifferent:	$\mathbf{IN}\phi \stackrel{\text{def}}{=} \mathbf{N}\mathrm{U}\phi \;\&\; \mathbf{N}\mathrm{U}\neg\phi$
Offence of Omission:	$\mathbf{OF}\neg\phi \stackrel{\text{def}}{=} \mathbf{N}\mathrm{U}\phi \;\&\; \mathbf{B}\mathrm{D}\neg\phi$
Impermissible:	$\mathbf{IM}\phi \stackrel{\text{def}}{=} \mathbf{B}\mathrm{D}\phi \;\&\; \mathbf{G}\mathrm{D}\neg\phi$
Offence of Commission:	$\mathbf{OF}\phi \stackrel{\text{def}}{=} \mathbf{B}\mathrm{D}\phi \;\&\; \mathbf{N}\mathrm{U}\neg\phi$
Totally Offensive:	$\mathbf{TO}\phi \stackrel{\text{def}}{=} \mathbf{B}\mathrm{D}\phi \;\&\; \mathbf{B}\mathrm{D}\neg\phi$

Chisholm also identifies the *permissible* with what is not forbidden ([Chisholm, 1963b], p.11), and so derivatively:

$$\text{(CPE)} \quad \mathbf{PE}\phi \stackrel{\text{def}}{=} \neg\mathbf{B}\mathrm{D}\phi \vee \neg\mathbf{G}\mathrm{D}\neg\phi$$

Exploring what Chisholm needs for a logic of good and bad Since Chisholm has already said that he intends *being good* and *being bad* as

contraries, let's encode this as "Chisholm's Contrariety Thesis":

(CCT) $\vdash \neg(\mathbf{G}\mathrm{D}\phi \;\&\; \mathbf{B}\mathrm{D}\phi)$[36]

Let us also add Chisholm's negative point stressed above about the application of "good" and "bad": that "their application to performance is logically independent of their application to non-performance" [Chisholm, 1963b, p. 10]. Call this the "Chisholm Independence Constraint":

(CIC) $\nvdash (\mathbf{G}\mathrm{D}\phi \rightarrow \mathbf{B}\mathrm{D}\neg\phi), \nvdash (\mathbf{B}\mathrm{D}\neg\phi \rightarrow \mathbf{G}\mathrm{D}\phi), \nvdash (\mathbf{G}\mathrm{D}\neg\phi \rightarrow \mathbf{B}\mathrm{D}\phi)$, and $\nvdash (\mathbf{B}\mathrm{D}\phi \rightarrow \mathbf{G}\mathrm{D}\neg\phi)$[37]

CIC is needed, since if what it says are not theses were in fact all theses, not only would redundancies result, but some categories would be incoherent. On the first point, the definitions of the *obligatory* and the *forbidden* would be redundant in one conjunct, since they would now be equivalent respectively to $\mathbf{G}\mathrm{D}\phi$ and to $\mathbf{B}\mathrm{D}\phi$; similar for the indifferent. On the second and more important point, the four categories of supererogatory commissions and omissions and of offensive commissions and omissions would be rendered incoherent, as the reader can easily check (e.g. $\mathbf{S}\mathrm{U}\phi$, by definition, entails $\mathbf{G}\mathrm{D}\phi \;\&\; \neg\mathbf{B}\mathrm{D}\neg\phi$).

However, Chisholm really needs something stronger than CIC, as illustrated by a natural alternative reading of "their application to performance is logically independent of their application to non-performance", namely:

(CIC′) $\nvdash (\mathbf{G}\mathrm{D}\phi \rightarrow \neg\mathbf{G}\mathrm{D}\neg\phi)$ and $\nvdash (\mathbf{B}\mathrm{D}\phi \rightarrow \neg\mathbf{B}\mathrm{D}\neg\phi)$

CIC′ is needed by Chisholm because it makes the intended logical space for the *totally offensive* and for the *totally supererogatory*[38], whereas CIC is not sufficient for that:

Totally Supererogatory ($\mathbf{T}$S)	$\mathbf{T}\mathrm{S}\phi \stackrel{\text{def}}{=} \mathbf{G}\mathrm{D}\phi \;\&\; \mathbf{G}\mathrm{D}\neg\phi$
Totally Offensive ($\mathbf{T}$O):	$\mathbf{T}\mathrm{O}\phi \stackrel{\text{def}}{=} \mathbf{B}\mathrm{D}\phi \;\&\; \mathbf{B}\mathrm{D}\neg\phi$

[36]For now, this expresses a *desired* status — to be a thesis; below we will specify a logic for which it is an axiom.

[37]With the addition of RE principles in a moment, the first two disjuncts suffice to cover the intended independence.

[38]We return to the merits of including them again below.

Furthermore, if the theses that are rejected in CIC actually held, then so too would those rejected in CIC′, given CCT (e.g. per CCT, if $\vdash (\mathbf{G}\mathrm{D}\phi \rightarrow \mathbf{B}\mathrm{D}\neg\phi)$ then $\vdash(\mathbf{G}\mathrm{D}\phi \rightarrow \neg\mathbf{G}\mathrm{D}\neg\phi))$[39] More specifically, if any of the four cited formulae in CIC were theses, then at least one of the two formulas cited in CIC′ would be a thesis as well. Thus CIC′ entails CIC (given double negation replacement for **G**D and **B**D and CCT), but CIC does not entail CIC′ (as is easily shown with the semantics to be offered below). CIC′ is thus the stronger independence claim, and so let this be the revised independence constraint needed for Chisholm's intended scheme.

Chisholm cites the following in passing as consequences of his scheme:

$\mathbf{O}\mathrm{B}\phi$ iff $\mathbf{I}\mathrm{M}\neg\phi$
$\mathbf{I}\mathrm{M}\phi$ iff $\mathbf{O}\mathrm{B}\neg\phi$
$\mathbf{P}\mathrm{E}\phi$ iff $\neg\mathbf{I}\mathrm{M}\phi$
$\mathbf{I}\mathrm{N}\phi \rightarrow \neg(\mathbf{S}\mathrm{U}\phi \vee \mathbf{S}\mathrm{U}\neg\phi \vee \mathbf{O}\mathrm{F}\phi \vee \mathbf{O}\mathrm{F}\neg\phi)$
$\mathbf{O}\mathrm{F}\phi \rightarrow \mathbf{P}\mathrm{E}\phi / \mathbf{O}\mathrm{F}\neg\phi \rightarrow \mathbf{P}\mathrm{E}\phi$
$\mathbf{S}\mathrm{U}\phi \rightarrow \neg\ \mathrm{OB}\phi / \mathbf{S}\mathrm{U}\neg\phi \rightarrow \neg\mathbf{O}\mathrm{B}\neg\phi$

Note that this reveals that Chisholm is assuming that **G**D and **B**D satisfy something like replacement of logical equivalents (at least for double negation). For although the last four follow from the definitions by PL alone, the first two do not, for without an allowance for substituting "$\neg\neg\phi$" for "ϕ" in the scope of "**G**D" and of "**B**D", these would not be provable (e.g. the second amounts to $\mathbf{G}\mathrm{D}\phi$ & $\mathbf{B}\mathrm{D}\neg\phi) \leftrightarrow (\mathbf{B}\mathrm{D}\neg\phi$ & $\mathbf{G}\mathrm{D}\neg\neg\phi)$. So let's add two RE principles and assume that GD and BD are classical modal operators:

(GD-RE) If $\vdash \phi \leftrightarrow \psi$, then $\vdash \mathbf{G}\mathrm{D}\phi \leftrightarrow \mathbf{G}\mathrm{D}\psi$
(BD-RE) If $\vdash \phi \leftrightarrow \psi$, then $\vdash \mathbf{B}\mathrm{D}\phi \leftrightarrow \mathbf{B}\mathrm{D}\psi$

There are other *desirable* consequences not cited by Chisholm, for example, where $\mathbf{O}\mathrm{P}\phi \stackrel{\mathrm{def}}{=} \mathbf{P}\mathrm{E}\phi$ & $\mathbf{P}\mathrm{E}\neg\phi$ (i.e. $(\neg\mathbf{B}\mathrm{D}\phi \vee \neg\mathbf{G}\mathrm{D}\neg\phi)$ & $(\neg\mathbf{B}\mathrm{D}\neg\phi \vee \neg\mathbf{G}\mathrm{D}\neg\phi)$, an important expansion of Urmson's Criterion follows, namely that anything that is supererogatory *or an offence* is such that it is optional but not indifferent:

(UC′) $(\mathbf{S}\mathrm{U}\phi \vee \mathbf{O}\mathrm{F}\phi) \rightarrow (\mathbf{O}\mathrm{P}\phi$ & $\neg\mathbf{I}\mathrm{N}\phi)$

For UC′ follows by definition from PL and RE. Also, with (but not with-

[39] Assuming replacement of equivalents.

out) RE principles, the desirable indifference of indifference to negation principle follows as well:

$$\mathbf{I}\mathrm{N}\phi \leftrightarrow \mathbf{I}\mathrm{N}\neg\phi$$

Likewise for the totally offensive and the totally supererogatory:

$$\mathbf{T}\mathrm{O}\phi \leftrightarrow \mathbf{T}\mathrm{O}\neg\phi$$
$$\mathbf{T}\mathrm{S}\phi \leftrightarrow \mathbf{T}\mathrm{S}\neg\phi$$

The following principle of "No Supererogatory Conflicts",

(NSC) $\neg(\mathbf{S}\mathrm{U}\phi \;\&\; \mathbf{S}\mathrm{U}\neg\phi)$

is derivable ($\mathbf{S}\mathrm{U}\phi$ entails $\mathbf{G}\mathrm{D}\phi$, but $\mathbf{S}\mathrm{U}\neg\phi$ entails $\neg\mathbf{G}\mathrm{D}\phi$), and we argued in 2.2 that it is a plausible constraint for this target concept. "No *Offence* Conflicts" are ruled out in the same way:

(NOC) $\neg(\mathbf{O}\mathrm{F}\phi \;\&\; \mathbf{O}\mathrm{F}\neg\phi)$

Furthermore, Meinong's plausible *Deontic-Axiological* bridge principles are derivable as well:

(MD-A) $(\mathbf{S}\mathrm{U}\phi \vee \mathbf{O}\mathrm{B}\phi \rightarrow \mathbf{G}\mathrm{D}\phi$
$(\mathbf{O}\mathrm{F}\phi \vee \mathbf{I}\mathrm{M}\phi) \rightarrow \mathbf{B}\mathrm{D}\phi$

Thus with CCT, and **O**B-RE, given our recasting of Chisholm's definitional scheme, we can generate a number of plausible and desirable consequences. In fact, really the main categorical contention expected is indeed provable at this point.

Chisholm's categorical scheme is indeed a partition Perhaps the most important thesis of the scheme in the context of the article would be that the nine categories above form a genuine partition — they are mutually exclusive and jointly exhaustive as reflected in Figure 11 (intended to exhaust the domain of Jane Doe's performances and non-performances).

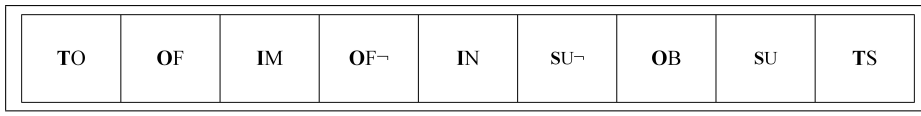


Figure 11: Chisholm's ninefold classification

Symbolically, the Chisholm Ninefold Classification is:

(CNC) $\mathrm{MJ}^9(\mathbf{T}\mathrm{O}\phi,\ \mathbf{O}\mathrm{F}\phi,\ \mathbf{I}\mathrm{M}\phi,\ \mathbf{O}\mathrm{F}\neg\phi,\ \mathbf{I}\mathrm{N}\phi,\ \mathbf{S}\mathrm{U}\neg\phi,\ \mathbf{O}\mathrm{B}\phi,\ \mathbf{S}\mathrm{U}\phi,\ \mathbf{T}\mathrm{S}\phi)$

Chisholm does not offer any proof that he has articulated an exhaustive and mutually exclusive classification, but a proof in our regimentation of Chisholm's scheme is straightforward, though tedious and left aside here.

Note that by implication, Chisholm is here rejecting Meinong's Fivefold classification:

(M5FC) $\mathrm{MJ}^5(\mathbf{S}\mathrm{U}\phi, \mathbf{O}\mathrm{B}\phi, \mathbf{I}\mathrm{N}\phi, \mathbf{O}\mathrm{F}\phi, \mathbf{I}\mathrm{M}\phi)$

Although mutual exclusiveness of these five categories is retained, Chisholm must reject the exhaustiveness implication. For the following is derivable from what we have already:

$\vdash\ (\mathbf{S}\mathrm{U}\neg\phi \vee \mathbf{T}\mathrm{O}\phi \vee \mathbf{T}\mathrm{S}\phi \vee \mathbf{O}\mathrm{F}\neg\phi) \rightarrow \neg(\mathbf{S}\mathrm{U}\phi \vee \mathbf{O}\mathrm{B}\phi \vee \mathbf{I}\mathrm{N}\phi \vee \mathbf{O}\mathrm{F}\phi \vee \mathbf{I}\mathrm{M}\phi)$

Obviously, a partition at the level of GD and BD follows, as Chisholm notes in passing (Figure 12).

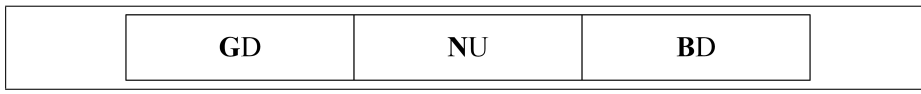


Figure 12: Chisholm's threefold axiological classification

Call this "Chisholm's Threefold Axiological Classification":

(CTAC) $\mathrm{MJ}^3(\mathbf{G}\mathrm{D}, \mathbf{B}\mathrm{D}, \mathbf{N}\mathrm{U})$

Examination of the definitions of the categories reveals the following association of **B**D, **G**D, and **N**U with the nine defined categories as in Figure 13.

Let me pause here to draw out what I take to be an important presupposition about the conceptualization of the two axiological operators vis

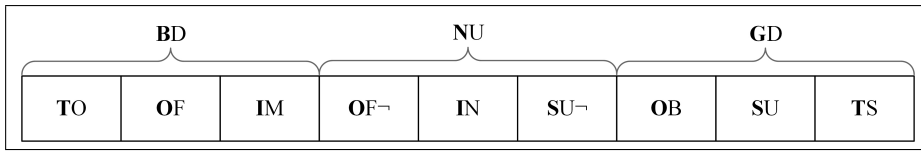


Figure 13: Chisholm's nine-fold classification with axiological triad

a vis that of the deontic notion of obligation involved. Note that the following is plainly entailed:

(NC) $\mathbf{O}\mathrm{B}\phi \rightarrow \neg\mathbf{O}\mathrm{B}\neg\phi$.[40]

Furthermore, Chisholm surely recognized this. But I argued earlier in Section 1.2 that endorsing NC amounts to viewing **O**B as expressing a notion of what is *overridingly* obligatory: obligatory and such that it trumps any other obligations that conflict with it. NC for **O**B suggests that first, for coherence, **G**D and **B**D would need to be read *at least as strongly as* "it is all things considered good that" and "it is all things considered bad that", respectively. More importantly, the central thesis CCT, $\neg(\mathbf{G}\mathrm{D}\phi \;\&\; \mathbf{B}\mathrm{D}\phi)$, would not be plausible at all on weak readings of **G**D and **B**D, for what could rule out something's being both good in some respects and bad in some other respects? So Chisholm must be presupposing *at least* these:

GDϕ only if it is good *all things considered* that ϕ.
BDϕ only if it is bad *all things considered* that ϕ.[41]

More could be said here, but we leave this for another occasion.

A deduction system: RCGB^1 (reconstruction of Chisholm's good-bad logic). Looking backwards, and continuing to cast things via operators, it appears that the *minimal* logic needed to generate the main claims Chisholm makes about his favored scheme is the following "*Reconstructed Chisholm Good-Bad Logic*":

[40]By definition, it amounts to $(\mathbf{G}\mathrm{D}\phi \;\&\; \mathbf{B}\mathrm{D}\neg\phi) \rightarrow \neg(\mathbf{G}\mathrm{D}\neg\phi \;\&\; \mathbf{B}\mathrm{D}\neg\neg\phi)$, and given RE and CCT, $\mathbf{G}\mathrm{D}\phi$ rules out $\mathbf{B}\mathrm{D}\neg\neg\phi$, and $\mathbf{B}\mathrm{D}\neg\phi$ rules out $\mathbf{G}\mathrm{D}\neg\phi$, either one of which suffices.

[41]The work of Chisholm-Sosa also suggests agreement with this interpretation of *good* and *bad*.

(RCGB^1)	(Taut)	All tautologies
	(MP)	If $\vdash \phi$ and $\vdash \phi \rightarrow \psi$, then $\vdash \psi$
	(CCT)	$\vdash \neg(\mathbf{G}\mathrm{D}\phi \ \& \ \mathbf{B}\mathrm{D}\phi)$
	(**G**D-RE)	If $\vdash \phi \leftrightarrow \psi$ then $\vdash \mathbf{G}\mathrm{D}\phi \leftrightarrow \mathbf{G}\mathrm{D}\psi$
	(**B**D-RE)	If $\vdash \phi \leftrightarrow \psi$ then $\vdash \mathbf{B}\mathrm{D}\phi \leftrightarrow \mathbf{B}\mathrm{D}\psi$

As mentioned, nothing is said by Chisholm to guide us on the behavior of "**G**D" and "**B**D" in the scope of action compounds, and so not on truth-functional connectives in our reconstruction either.

Let me quickly note some derivative principles governing the usual operators of SDL, but we must forgo careful comparison for another time. Recall the definition of **O**B:

$$\mathbf{O}\mathrm{B}\phi \stackrel{\text{def}}{=} \mathbf{G}\mathrm{D}\phi \ \& \ \mathbf{B}\mathrm{D}\neg\phi$$

Replacement of provable equivalents for **O**B,

(**O**B-RE) If $\vdash \phi \leftrightarrow \psi$, then $\vdash \mathbf{O}\mathrm{B}\phi \leftrightarrow \mathbf{O}\mathrm{B}\psi$,

follows immediately from **B**D-RE and **G**D-RE and the definition of **O**B,

as does the duality of **O**B and **P**E, for example

$$\vdash \mathbf{P}\mathrm{E}\phi \leftrightarrow \neg\mathbf{O}\mathrm{B}\neg\phi$$

Similarly for other standard deontic inter-definability equivalences involving **O**B, **P**E, and **I**M when construed on this scheme. As we saw above in discussing the presupposed readings of **G**D, **B**D, and **O**B, the characteristic D axiom for SDL is derivable and presupposed:

(NC) $\vdash \mathbf{O}\mathrm{B}\phi \rightarrow \neg\mathbf{O}\mathrm{B}\neg\phi$

Indeed, something slightly stronger than what is needed is derivable:

$$\vdash \mathbf{O}\mathrm{B}\phi \rightarrow (\neg\mathbf{B}\mathrm{D}\phi \ \& \ \neg\mathbf{G}\mathrm{D}\neg\phi),$$

for only *the disjunction* of the consequent's disjuncts is needed to generate NC.

However, we have no guidance on the logical behavior of **G**D and **B**D applied to the *verum*, *falsum*, or to compounds like conjunctions, disjunctions, or material implications. All we seem to be able to con-

clude for **O**B is just this derivable fragment of SDL, essentially just the classical system ED again (what I called "The Traditional Scheme":

(**O**B-RE) If $\vdash \phi \leftrightarrow \psi$, then $\vdash \mathbf{O}\mathrm{B}\phi \leftrightarrow \mathbf{O}\mathrm{B}\psi$
(NC) $\vdash \mathbf{O}\mathrm{B}\phi \rightarrow \neg\mathbf{O}\mathrm{B}\neg\phi$
(Def-PE) $\vdash \mathbf{P}\mathrm{E}\phi \leftrightarrow \neg\mathbf{O}\mathrm{B}\neg\phi$

In particular, there is no carry over to the behavior of **O**B characteristic of SDL axioms like our earlier (A2), (A3), and (A4) from Section 1.2:

(A2) $(\mathbf{O}\mathrm{B}\phi \ \& \ \mathbf{O}\mathrm{B}\psi) \rightarrow \mathbf{O}\mathrm{B}(\phi \ \& \ \psi)$ [a C principle]
(A3) $\mathbf{O}\mathrm{B}(\phi \ \& \ \psi) \rightarrow (\mathbf{O}\mathrm{B}\phi \ \& \ \mathbf{O}\mathrm{B}\psi)$ [an M principle]
(A4) $\mathbf{O}\mathrm{B}\top$ [an N principle].

We must postpone for elsewhere to explore what it would take to generate each of (A2)-(A4), and whether or not there might be reasons to not do so implicit in Chisholm's work that might shed light on some of the classical puzzles for SDL.[42] However, in the next section we consider one more principle of SDL (and of many weaker systems), $\neg\mathbf{O}\mathrm{B}\bot$, which deserves separate consideration.

Two minor expansions of RCGB^1 and comparison to SDL. Note one oddity with the logic generated for **O**B: although $\mathbf{O}\mathrm{B}\phi \rightarrow \neg\mathbf{O}\mathrm{B}\neg\phi$ is a theorem, this *intuitively weaker* and more widely endorsed principle is *not* a theorem:

(OD) $\neg\mathbf{O}\mathrm{B}\bot$[43]

We thus have the oddity that although there can be no conflicts of obligation, there can be contradictory obligations. Here the two thesis, NC and OD are distinguished (unlike in SDL), but the distinction, *as made*, is unwelcome. For the most plausible stories that would make room for the possibility of contradictory obligations (e.g. I solemnly promise you on your death bed to complete your lifelong quest to square the circle) will plausibly be expected to generate obligations that conflict with one another as well (I promise each of two contingent but mutually incompatible things to you on your deathbed, perhaps unwittingly).

Furthermore, the plausibility of OD seems to be retained when read via its definiens and an application of RE:

[42]Similarly for the exploration of mixed axiological-deontic formula, having only noted one or two above.

[43]See [Chellas, 1980] and [Schotch and Jennings, 1981] for early background.

(OD′) $\neg\mathbf{G}\mathrm{D}\bot \vee \neg\mathbf{B}\mathrm{D}\top$

It seems plausible that a contradictory state of affairs is not all in all good, and that a tautologous state of affairs is not all in all bad, and these claims fit the spirit of Chisholm's scheme. Firstly, a special instance of CCT is $\mathbf{G}\mathrm{D}\bot \rightarrow \neg\mathbf{B}\mathrm{D}\bot$. But it seems that the only thing that could be said in support of the possibility that $\mathbf{G}\mathrm{D}\bot$ is that a contradictory state of affairs is all in all good *and* all in all bad; but CCT rules that out as *axiologically* false. Secondly, it is simply hard to imagine anything that could be said in support of the idea that a tautologous condition could possibly be *overall bad*.
So I think it is plausible to endorse OD' as befitting this recast of Chisholm, and indeed, the suggestion is for something stronger than needed:

(OD″) $\neg\mathbf{G}\mathrm{D}\bot \;\&\; \neg\mathbf{B}\mathrm{D}\top$

We might accordingly consider two natural expansions of the base logic RCGB^1:

(RCGB^2) RCGB^1 + OD′
(RCGB^3) RCGB^1 + OD″

RCGB^2 is all that is required to derive OD, and with either expansion of RCGB^1, we can derive an expanded fragment of SDL.

RCGB^2's derivable fragment of SDL:	
(**O**B-RE)	If $\vdash \phi \leftrightarrow \psi$, then $\vdash \mathbf{O}\mathrm{B}\phi \leftrightarrow \mathbf{O}\mathrm{B}\psi$,
(NC)	$\vdash \mathbf{O}\mathrm{B}\phi \rightarrow \neg\mathbf{O}\mathrm{B}\neg\phi$.
(Def-**P**E)	$\vdash \mathbf{P}\mathrm{E}\phi \leftrightarrow \neg\mathbf{O}\mathrm{B}\neg\phi$
(OD)	$\vdash \neg\mathbf{O}\mathrm{B}\bot$

Semantics for RCGB logics

Chisholm's writing on deontic logic did not involve any formal semantics, but we offer a simple one here using "neighborhood semantics" [Chellas, 1980]. A frame, (W, G, B), contains a set of worlds, W, and two functions that map worlds to sets of sets of worlds (often thought of as sets of propositions):

(G) $G : W \to \text{Pow}(\text{Pow}(W))$, i.e. $G(u) \subseteq Pow(W)$
(B) $B : W \to \text{Pow}(\text{Pow}(W))$, i.e. $B(u) \subseteq Pow(W)$

So the value of the goodness function for any given world, u, is a set of subsets of W — the propositions the function assigns as good per u. Similarly, for the badness function. To get a model, M, we add a valuation function, v, assigning sets of worlds to the atomic sentences, extended in the usual way for the truth-functional compounds. The truth conditions (relative to a model) for the goodness and badness operators are:

[**G**D] $M, u \vDash \mathbf{G}\text{D}\phi$ iff $\| \phi \| \in G(u)$
[**B**D] $M, u \vDash \mathbf{B}\text{D}\phi$ iff $\| \phi \| \in B(u)$

We then add an additional constraint on the frames, namely that they must validate CCT:

(CCCT) $B(u) \cap G(u) = \varnothing$, for any $u \in W$, in any frame

All the formulae said above to be non-derivable can be easily shown to be invalid by constructing counter-models using our RCGB semantic framework, and thereby shown to be non-derivable as well in the corresponding logic (since the three RCGB logics are sound). For example, suppose $W = \{i\}$, and $G(i) = B(i) = \varnothing$. Then for any model based on this, $\mathbf{G}\text{D}\top$ and $\mathbf{B}\text{D}\bot$ are each false at i; this also shows that a necessity rule for **G**D (from $\vdash \phi$, derive $\vdash \mathbf{G}\text{D}\phi$) is not validity-preserving (nor derivable), and that $\mathbf{O}\text{B}\top$ is invalid as well in the RCGB[1] logic.

Truth Conditions for the remaining operators are easily derived:

(**T**O)	M, $u \vDash \mathbf{T}\text{O}\phi$ iff	$\| \phi \| \in B(u)$ & $\| \neg\phi \| \in B(u)$
(**O**F)	M, $u \vDash \mathbf{O}\text{F}\phi$ iff	$\| \phi \| \in B(u)$ & $\| \neg\phi \| \notin G(u)$ & $\| \neg\phi \| \notin B(u)$
(**I**M)	M, $u \vDash \mathbf{I}\text{M}\phi$ iff	$\| \phi \| \in B(u)$ & $\| \neg\phi \| \in G(u)$
(**I**N)	M, $u \vDash \mathbf{I}\text{N}\phi$ iff	$\| \phi \| \notin G(u)$ & $\| \phi \| \notin B(u)$ & $\| \neg\phi \| \notin G(u)$ & $\| \neg\phi \| \notin B(u)$
(**O**B)	M, $u \vDash \mathbf{O}\text{B}\phi$ iff	$\| \phi \| \in G(u)$ & $\| \neg\phi \| \in B(u)$
(**S**U)	M, $u \vDash \mathbf{S}\text{U}\phi$ iff	$\| \phi \| \in G(u)$ & $\| \neg\phi \| \notin B(u)$ & $\| \neg\phi \| \notin G(u)$
(**T**S)	M, $u \vDash \mathbf{T}\text{S}\phi$ iff	$\| \phi \| \in G(u)$ & $\| \neg\phi \| \in G(u)$
(**N**Uϕ)	M, $u \vDash \mathbf{O}\text{P}\phi$ iff	$\| \phi \| \notin G(u)$ & $\| \phi \| \notin B(u)$
(**P**Eϕ)	M, $u \vDash \mathbf{P}\text{E}\phi$ iff	$\| \phi \| \notin B(u)$ or $\| \neg\phi \| \notin G(u)$
(**O**Pϕ)	M, $u \vDash \mathbf{O}\text{P}\phi$ iff	($\| \phi \| \notin B(u)$ or $\| \neg\phi \| \notin G(u)$) & ($\| \neg\phi \| \notin B(u)$ or $\| \phi \| \notin G(u)$)

As with GD and BD, the various claims about underivability for various principles governing OB and other non-derivability claims about relationships between the defined operators (like Meinong's first two Laws of Omission for Supererogation and Offence) are easily confirmed using this semantics.

The frame conditions needed to validate OD′ and OD″ are, respectively:

(COD′) Either $\varnothing \notin G(u)$ or $W \notin B(u)$, for each $u \in W$,
(COD″) Both $\varnothing \notin G(u)$ and $W \notin B(u)$, for each $u \in W$.

Let's define three classes of models:

RCGB^1 Models: All RCGB models
RCGB^2 Models: All RCGB models where COD′ holds
RCGB^3 Models: All RCGB models where COD″ holds

The following are easily shown:

Metatheorems: 1) RCGB^1 is determined by the RCGB^1 models.
2) RCGB^2 is determined by the RCGB^2 models.
3) RCGB^3 is determined by the RCGB^3 models.

Reflections on Chisholm's main scheme

Doubts about the totally supererogatory and the totally offensive

Recall the two categories expressed by **T**S and **T**O:

Totally Offensive (**T**O): $\mathbf{T}\mathrm{O}\phi \stackrel{\text{def}}{=} \mathbf{B}\mathrm{D}\phi \;\&\; \mathbf{B}\mathrm{D}\neg\phi$
Totally Supererogatory (**T**S): $\mathbf{T}\mathrm{S}\phi \stackrel{\text{def}}{=} \mathbf{G}\mathrm{D}\phi \;\&\; \mathbf{G}\mathrm{D}\neg\phi$

The choice of labels is odd, since nothing "totally supererogatory" is supererogatory and nothing "totally offensive" is offensive — the adjectives are not detachable from the adverb-adjective labels, as one might expect (e.g. as with "totally exhausted" and "exhausted").[44] Setting the issue of the odd labels aside, Chisholm suggests the defined conditions are needed and credibly satisfiable by situations acknowledged as realizable by possible ethical theories he wants his scheme to be able to

[44]Chisholm indicates he is inspired in his terminological choice by analogy with the use of "'totally' ... in the theory of relations".

accommodate. Nonetheless, they have been viewed subsequently with much suspicion, and unlike the other categories, they have not been taken up by others.[45]

Ruling out the categories of the totally supererogatory and the totally offensive is straightforward — it amounts to just endorsing axiological *no conflict-principles* of the sort mentioned in CIC' for each of the basic operators taken individually:

(NC-GD) $\mathbf{G}\mathrm{D}\phi \rightarrow \neg\mathbf{G}\mathrm{D}\neg\phi$
(NC-BD) $\mathbf{B}\mathrm{D}\phi \rightarrow \neg\mathbf{B}\mathrm{D}\neg\phi$

Given the preceding reflections tending to favor the rejection of the totally supererogatory and totally offensive, let's introduce three further possible expansions of RCGB^1, and in this case, expansions that are more substantive in that they result in contractions of the possible normative positions in Chisholm's scheme.

$\mathrm{RCGB}^4 = \mathrm{RCGB}^1 + \text{NC-}\mathbf{G}\mathrm{D}$
$\mathrm{RCGB}^5 = \mathrm{RCGB}^1 + \text{NC-}\mathbf{B}\mathrm{D}$
$\mathrm{RCGB}^6 = \mathrm{RCGB}^1 + \text{NC-}\mathbf{G}\mathrm{D} + \text{NC-}\mathbf{B}\mathrm{D}$

Among other things, adding these axiological no-conflicts principles reduces the partition above to a Chisholm-like seven-fold classification of alternatives:

(CSC) $\quad \mathrm{MJ}^7(\mathbf{O}\mathrm{F}\phi, \mathbf{I}\mathrm{M}\phi, \mathbf{O}\mathrm{F}\neg\phi, \mathbf{I}\mathrm{N}\phi, \mathbf{S}\mathrm{U}\neg\phi, \mathbf{O}\mathrm{B}\phi, \mathbf{S}\mathrm{U}\phi)$

Figure 14 provides a diagrammatic expression of C7FC.

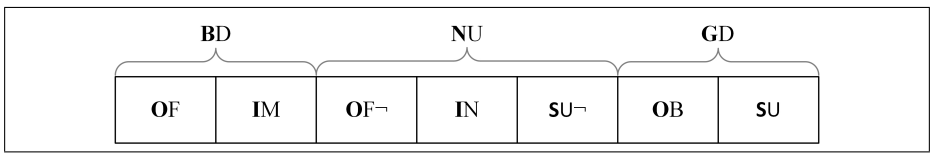


Figure 14: Chisholm's seven-fold classification

Note that [Chisholm and Sosa, 1966b] develops a framework for *intrinsic preferability*, and in [Chisholm and Sosa, 1966a] they apply it to the problem of supererogation. Of particular interest here is that in both papers an analysis of *good* and *bad* is given, and then a categorical

[45] Also, in the work by Chisholm-Sosa, these two of his nine categories here are ruled out as impossible.

scheme is developed in which there are just *the seven categories above*. The conditions definitive of the totally supererogatory and of the totally offensive are not merely missing, but their defining conditions are ruled out as logically impossible by the framework.

Given these axiological *no conflict-principles*, along with **G**D-RE and **B**D-RE, the prior definitions for offensive and supererogatory commissions and omissions could also be simplified to the following (by eliminating now-redundant clauses):

Offence of Commission (**O**F)	$\mathbf{O}\mathrm{F}\phi \stackrel{\text{def}}{=} \mathbf{B}\mathrm{D}\phi \ \& \ \neg\mathbf{G}\mathrm{D}\neg\phi$
Offence of Omission (**O**F$\neg$)	$\mathbf{O}\mathrm{F}\neg\phi \stackrel{\text{def}}{=} \mathbf{B}\mathrm{D}\neg\phi \ \& \ \neg\mathbf{G}\mathrm{D}\phi$
Supererogatory Omission (**S**U$\neg$)	$\mathbf{S}\mathrm{U}\neg\phi \stackrel{\text{def}}{=} \mathbf{G}\mathrm{D}\neg\phi \ \& \ \neg\mathbf{B}\mathrm{D}\phi$
Supererogatory Commission (**S**U)	$\mathbf{S}\mathrm{U}\phi \stackrel{\text{def}}{=} \mathbf{G}\mathrm{D}\phi \ \& \ \neg\mathbf{B}\mathrm{D}\neg\phi$

Lastly, given the semantic framework above, to validate these NC principles, we would just add the following semantic clauses:

(NC-G)	For any u in W and $X \subseteq W$, $X \in G(u)$ only if $W - X \notin G(u)$
(NC-B)	For any u in W and $X \subseteq W$, $X \in B(u)$ only if $W - X \notin B(u)$

These determine the three additional RCGB logics:

Metatheorems:
(4) RCGB^4 is determined by the RCGB models where NC-G holds.
(5) RCGB^5 is determined by the RCGB models where NC-B holds.
(6) RCGB^6 is determined by the RCGB models where both hold.

Let's turn to some other possible objections and then move on.

Additional objections to Chisholm's main scheme including adequacy of analyses

a) The need for constraints on the interpretation of *good* and *bad*: Some constraints seem to be needed on the interpretation of "good" and "bad". As already noted, for them to play the role they do both at the axiological level, given CCT, and at the level of the defined terms, esp. **O**B, it seems we must interpret **G**D and **B**D as expressing at least *all things considered* notions. But it would seem that further constraint is needed. If we interpret "good" and "bad" without any qualification, various implausible consequences

arise at the level of the defined moral notions, since prudential issues will turn into moral ones automatically. Given the set of target concepts being defined, it seems we must read the axiological notions as focusing on some *morally relevant sense* if they will generate something like the moral notions targeted by Chisholm. Furthermore, we must also assume that the morally relevant notions of moral goodness and of moral badness do not involve any of the notions being defined (e.g. permissibility) by them, else circularity of analysis sets in.

b) Problem revealed in the analysis of supererogation:
A more serious problem can be raised about the adequacy of the analysis of supererogation. In arguing against attempts to subsume supererogation under the concept of *imperfect obligations*, he seems to overlook a problematical implication of his own view. He imagines a case where I must give surplus goods to someone in a group of individuals, but no one in particular, and he asserts that:

> "In giving to Jones, I do, *ipso facto*, give to Jones or to Robinson ... or to Smith; hence I do fulfill my entire obligation, and it would be incorrect, therefore, to suppose that the act is a case of 'non-obligatory well-doing'." [Chisholm, 1963b, p. 4]

It seems Chisholm wants to say giving the goods to Jones is good *and obligatory*, and so not supererogatory. This is puzzling. First, how can it *not* be *non*-obligatory to give to Jones, since by stipulation I could fulfill the only obligation in focus by giving to one of the others instead? That giving to Jones suffices to fulfill my obligation completely does not entail that doing so is obligatory, and the description of the case itself rules out its being obligatory. Secondly, although he is right that it is not supererogatory in the case imagined, *does that result hold per his analysis of supererogation*? For convenience, assume it is money, say "$25", not goods that are to be given. Now suppose first that I fulfill the obligation by giving $25 to Jones. Since $\neg\mathbf{O}$Bj (j for *I give Jones $25*), on the proposed analysis, it follows that $\neg\mathbf{G}$Dj $\vee$ $\neg\mathbf{B}$D$\neg$j, and the latter disjunct is surely plausible in the imagined case. But it can plausibly be also good that I give Jones $25 (and it seems clear Chisholm is assuming that the act is good), even if not better to give it to him than any other, so suppose that too is the case. But

also notice that it is surely implausible in the imagined scenario that it is good to *not* give the $25 to Jones, since that can be accomplished by giving to no one, which is meant to be bad in the scenario Chisholm imagines — a failure to keep my obligation to give to at least one. But then it turns out to be *supererogatory* to give Jones $25, since the proposed defining condition for a supererogatory commission is met: $\mathbf{G}$Dj & $\neg\mathbf{G}$D$\neg$j & $\neg\mathbf{B}$D$\neg$j. Yet there appears to be nothing supererogatory about the imagined case (Chisholm is right about this aspect). I do not go *beyond* the call by selecting one of the possible beneficiaries and giving that one *a loan of the exact amount required* to be given to someone or other in the group. Contrast this with the cases where I give $50 to Jones or where I give $25 to Jones *and* $25 to Smith as well. Here I have surely met my obligation fully too, but I have also gone beyond the call in fulfilling that obligation (assuming I am not violating another obligation in the process of giving more).

So Chisholm's framework seems unable to distinguish the two sorts of cases. Some things that are good to do and neither good nor bad to skip are supererogatory, and some are not, so *the defining condition is not sufficient for supererogation.* At best, it looks like only the left to right implication holds and the other must be rejected:

$$\mathbf{S}\mathrm{U}\phi \rightarrow [\mathbf{G}\mathrm{D}\phi \;\&\; \neg(\mathbf{G}\mathrm{D}\neg\phi \vee \mathbf{B}\mathrm{D}\neg\phi)]$$
$$\mathbf{S}\mathrm{U}\phi \nleftarrow [\mathbf{G}\mathrm{D}\phi \;\&\; \neg(\mathbf{G}\mathrm{D}\neg\phi \vee \mathbf{B}\mathrm{D}\neg\phi)]$$

We will return to this sort of problem later and see that it has infested later accounts of supererogation, as noted by me, as well as Sven Ove Hansson [Hansson, 2013].

A secondary moral to be drawn from this case, specifically from the initial mistake Chisholm seems to make of thinking that because an action fulfills an obligation completely it follows that it is not supererogatory, is that it is important to reflect on supererogation and obligation in the context of agency, especially *dyadic agency* — bringing one thing about by bringing another about. For example, here one may fulfill one's obligation to give to one in the group by giving (exactly) $25 to Jones, the latter implying you have done the least you can do, but one may also fulfill it by giving Jones $50 and this has the status of being the fulfillment of a duty via a supererogatory pathway. We will take this up later.

c) Doubts about the analysis of obligation (and deontic kin): Also, in his discussion of offenses, Chisholm clearly asserts that things can be anywhere from slightly bad to very bad, yet not be impermissible, and things can be anywhere from slightly good to very good and yet not be obligatory. But then why exactly can't something be good (perhaps just somewhat all in all) to do and bad (perhaps just somewhat all in all) to not do, but not be obligatory? It might be somewhat good though time-absorbing to help someone lost asking for complex directions, and somewhat bad to not do so, but still be *optional*, since you are in a real rush and the cost to you might be almost as great as the cost to the stranger of waiting for someone else to help, so morality leaves it at your discretion as far as what it *demands*, even if it *recommends* that you help. (Cf. You ought to, but don't have to.) That is, there is reason to doubt that this is unsatisfiable,

$$\mathbf{O}\mathrm{P}\phi \;\&\; \mathbf{G}\mathrm{D}\phi \;\&\; \mathrm{BD}\neg\phi$$

especially in contexts like here, where we are assuming that morality leaves a fair amount of leeway, allowing one to do good that is not required, and to do bad that is not forbidden (e.g. good and ideal to not exercise some moral right you have, but somewhat bad all in all to do so, but not obligatory to not do so).[46] These reflections thus raise doubts that obligation can be defined via the condition, $\mathbf{G}\mathrm{D} \;\&\; \mathbf{B}\mathrm{D}\neg$, and so they raise doubts that any of the core deontic notions (obligation, permissibility, impermissibility) can be reduced in the manner indicated.[47]

d) **T**S, **T**O, and CCT: A potential instability in the foundation: Returning to the Totally Supererogatory and the Totally Offensive, if something might be all things considered good and its negation all things considered good as well, why is it that nothing can be all things considered good and all things considered bad? In a word, if

[46] These come closer to quasi-supererogation and quasi-offence, notions rightly stressed by Mellema ([Mellema, 1987], [Mellema, 1991]), and defined later on in Section 5.

[47] This ties in with what we said in the last sentence of objection a) above where we raised a different reason to wonder if the deontic notions could be defined without circularity in terms of *moral* goodness and badness, noting that *moral* goodness (badness), not just goodness (badness) per se, must be used if the aim is to analyze the target notions.

there can be intra-valent conflicts, why can't there be inter-valent conflicts? For example, if $\mathbf{G}\mathrm{D}\phi$ & $\mathbf{G}\mathrm{D}\neg\phi$ is satisfiable because ϕ is good all in all in various respects and $\neg\phi$ is good all in all in various respects, and neither of these good-producing respects outweighs the other, why couldn't $\mathbf{G}\mathrm{D}\phi$ & $\mathbf{B}\mathrm{D}\phi$ be satisfiable for similar reasons? CCT is needed for the definitional scheme he used to work, but it is not independently motivated.

The above objections are not meant to suggest that the notions of *good* and of *bad* might not have some important role in defining the notions of supererogation and offence, but they raise questions about whether they could suffice, and whether independent core deontic notions are not needed as well. This would mean, as admirable as the attempted analyses have been, and as illuminating and insightful as Chisholm's discussion of the subject has been in this seminal article, it is not clear that any of the target concepts can be adequately defined this way, and perhaps for reasons that have not fully registered on friends of supererogation either. Luckily for all of us, a lot can be learned from interesting efforts that don't ultimately succeed.[48]

We turn next to a rather different framework of Chisholm's that was influential in earlier developments of defeasibility reasoning, including in normative contexts, one that includes a rather strikingly different approach to supererogation.

4 Chisholm on the logic of requirement & supererogation and kin

Introduction The early to mid-1960s were a period of remarkably high activity for Chisholm in areas of interest to deontic logic (and elsewhere), and once again, in his work on the ethics/logic of requirement,

[48]Space limitations prevent us from exploring the joint Chisholm-Sosa work mentioned in passing above [Chisholm and Sosa, 1966a; Chisholm and Sosa, 1966b], but I do not think they escape the more substantive difficulties regarding the analysis of supererogation and the deontic notions mentioned above. For the definitions of those notions in terms of good and bad remain unchanged (although now themselves defined in terms of the notion of intrinsic preferability taken as basic in that framework). On the plus side, the *totally supererogatory* and *totally offensive* are logically impossible conditions and so not mentioned (despite references to [Chisholm, 1963b], and they also acknowledge that the notion of good and bad (and preferability) have to be qualified in some way to be used to capture the moral notions occurring in what we called above the reduced "Chisholm-like Sevenfold Classification".

we find him working at the cusp of informal deontic logic and ethical theory. However, his logic of requirement has been influential well beyond its potential applications in ethics and deontic logic (e.g. in epistemology), since many of its ideas are relevant to defeasible reasoning generally as well, he and John Pollock being early pioneers among philosophers in theorizing about defeasible reasoning. Our focus must be on the logic of requirement and the account of supererogation and kindred concepts that the framework might provide, but we will sketch and develop the underlying framework first.

In [Chisholm, 1964], his approach is, for the most part, that of conceptual analysis, with a focus on introducing a series of key definitions using a single dyadic primitive, with a smattering of symbolizations, given essentially parenthetically, with the definitions. There is no explicit representation using symbolic logic of axioms and theorems. The exposition in [Chisholm, 1974] is somewhat more formal, with some labeled definitions, specified axioms, and a few theorems listed, although it is less formal than the *intrinsic preferability* framework in [Chisholm and Sosa, 1966a]. I will continue to provide some modest regimentation, as well as often adding something naturally available in the conceptual framework, but not specified by Chisholm, and adding corrections where there is a mismatch in formulation and clear intention. I believe what I will present captures the view and its spirit.

[Chisholm, 1964; Chisholm, 1974] are influential primarily for their attempt to systematically analyze concepts that would later be in central focus in the deontic logic of defeasible normative notions (and their analogs in epistemology and AI on defeasible reasoning): *prima facie* duty, conflicts of obligation, defeated obligations, undefeated obligations, all things considered *oughts*, etc. Here I will exposit the framework, but with not much critical attention, in order to focus primarily on how supererogation and kin are weaved into this framework once developed.

I begin with the brief sketch of the 1964 article, which focuses on the core conceptual scheme, and then turn to a more regimented account in expositing the 1974 article. With the requirement framework articulated, I turn to the applications to supererogation and kindred notions, and finally to a comparison with the prior accounts that Chisholm endorsed, and an assessment of the theory, especially the applications. Note that I do not sketch a semantics suitable for Chisholm's framework here, as a fair amount of the logic needs to be developed as it is to reach the portions about supererogation and kin. I commend to the reader the work of Belzer and Loewer, which has strong affinities to Chisholm's

conceptual framework, and provides a semantics as well (although the object language does not employ propositional quantifiers). See also the work by Åqvist inspired by integrating themes from Chisholm's requirement work with tense logic, some of which does make use of propositional quantifiers.[49]

4.1 The 1964 account in the "Ethics of Requirement"[50]

"The Ethics of Requirement" (ER), [Chisholm, 1964], went into print in the year following his [Chisholm, 1963b]. Where the latter ended with a very bold claim, ER opens with one:

> By taking 'p requires q' as our single ethical primitive and making use of the concept of an *act*, we can define all the fundamental concepts of ethics. We can reduce a number of perplexing terms — e.g. 'good', 'obligatory', '*prima facie* duty', 'commitment', 'defeasibility', 'overrides', 'supererogatory', 'optional', 'indifferent' — to a single term which is not, in fact, restricted to ethics. [Chisholm, 1964, p. 147].[51]

Chisholm goes on to indicate that he thinks there are eight perplexing problems in practical reasoning that can be addressed fruitfully from this perspective, and in the second half of the paper (Sections 10-12, pp. 150-3), he addresses these *applications* briefly, focusing on our topic of supererogation and kin (Section 12, pp. 152-3) at greater length than the other seven problems. In the first half of the paper (Sections 2-9, pp. 147-50) he articulates the *general requirement framework*, and that will be the main focus in this Sub-Section (and the next).

As indicated in the quotation above, there will be one evaluative primitive, pRq, and although Chisholm speaks of *states of affairs and/or events* as relata and R as a relation, he nonetheless avails himself informally of an apparatus much like the quantified propositional logic that he and Sosa used explicitly in [Chisholm and Sosa, 1966a]. So I will reframe his approach more explicitly as quantifying over propositions, and adjust the readings accordingly with the thought that this meta-

[49] See for example, [Loewer and Belzer, 1983; Loewer and Belzer, 1991], [Belzer and Loewer, 1997]; and [Åqvist, 1985; Åqvist, 1993b; Åqvist, 1997b; Åqvist, 1998].

[50] [Chisholm, 1964].

[51] And the two bold claims may be related, since a key element in ER will be a proposed analysis of what ought to be, and what ought to be figures centrally in the brief sketch in the final paragraph of [Chisholm, 1963b].

physical issue in not central here, and Chisholm's own remarks accord with this.[52]

Chisholm illustrates what he has in mind by requirement with some examples, among which are: "promise-making requires–or calls for–promise-keeping, being virtuous, according to Kant, requires being rewarded; the dominant seventh requires the chord of the tonic; one color in the lower left calls for a complementary color in the upper right . . . " [Chisholm, 1964, p.147]

Chisholm then goes on to develop the conceptual/analytic framework, but with no formalities. He points out some things that should not hold given the intended interpretation of the requirement relation. He then begins using it to offer definitions/analyses of what it is for a proposition to be *de facto required*, what it is for one to be *overridden*, going on to famously indicate that "an overriding may itself be overridden" (p. 149), illustrating as follows:

$p\mathrm{R}q$
$(p \;\&\; r)\mathrm{R}\neg q$
$(p \;\&\; r \;\&\; s)\mathrm{R}q$
$(p \;\&\; r \;\&\; s \;\&\; t)\mathrm{R}\neg q$
. . .

He next introduces an analysis of what *ought to be* (clearly treating this as an *all things considered ought*), and then suggests an analysis of *good* and of *bad*, introduces an *agential construction (S brings it about that)*, gives his famous so-called *reduction of what one ought to do to what it ought to be that one does*, and suggests an analysis of *commitment*. We will develop most of these and more in the context of his more formal presentation of the same framework in 1974.

4.2 The 1974 account in "Practical Reason and the Logic of Requirement"[53]

Recasting and elaborating on the 1964 framework

He again takes R to be his primitive but he explicitly reads "$p\mathrm{R}q$" subjunctively as "p *would* require q", or as "p when it obtains requires q", or as "p is such that if it were to obtain it would require q" (p. 4). He also indicates that he will take the relata of R to stand in truth-functional

[52] Chisholm himself gives as an alternative reading of his definition of an *overridden requirement* one that begins with "there are true *propositions* p and s . . . " instead of "there are states of affairs p and s . . . " (p. 148).

[53] [Chisholm, 1974]

relations, and essentially, the theory is in fact layered over a classical propositional logic extended with propositional quantifiers, much like that used in [Chisholm and Sosa, 1966b]. There is a framework of nine definitions (we will weave in many others) that is much like the 1964 scheme (with minor differences such as the introduction of an alethic modal operator). But there is also a statement of basic principles presented as seven explicit axioms, and a list of five theorems (again we will list many more), along with five formulae listed as ones that should not be theorems.[54] A specification of the intended essential formulae might thus look like this (and we will refer to the logical framework to be sketched as "REQ"):

REQ formulas is the smallest set satisfying:
1) $p_1, \ldots, p_n$ are in REQ formulas;
2) If ϕ is in REQ formulas, then so is $\neg\phi$;
3) If ϕ and ψ are REQ formulas, so are $(\phi \vee \psi)$, $(\phi \,\&\, \psi)$, $(\phi \rightarrow \psi)$ and $(\phi \leftrightarrow \psi)$;
4) If ϕ is in REQ formulas, then so is $\forall \mathrm{v}\phi$, where v is a propositional variable;
5) If ϕ and ψ are in REQ formulas and contain no occurrences of R, then $(\phi \mathrm{R} \psi)$ is in REQ formulas;
6) If ϕ is in REQ formulas, so are $\Box\phi$ and $\mathbf{B}\mathrm{A}\phi$.[55]

Chisholm defines a variety of other terms, but we will represent these via abbreviational definitions and take the primitive language to contain just these formulae and primitive elements, which seems to best match Chisholm's presentation. Our focus will be on the spirit of Chisholm's framework as he articulates it, and its potential development, with only modest exploration of the logical properties of some of the various operators introduced, which are considerable, since the language is expressively powerful, especially given the propositional quantification.

Chisholm defines the de facto requirement relation, read simply as 'ϕ *requires* ψ', as follows:

(D1) $\quad \phi \mathrm{R}' \psi \stackrel{\text{def}}{=} \phi \,\&\, \phi \mathrm{R} \psi$. [$\phi$ (de facto) requires ψ]

So ϕ requires ψ iff ϕ is true and ϕ would require ψ. He explicitly en-

[54] There is no discussion of the background logic, no rules of inference, but it appears to be a mix of modal logic and extended propositional logic, and we will need to fill in in many places, but guided by the spirit of the framework.

[55] In what follows, where no confusion will result, we will often drop the outer parenthesis around formulae.

dorses a right and left RE rule for R:

(RE) If $\vdash \phi \leftrightarrow \psi$ then $\vdash \chi \mathrm{R} \phi \leftrightarrow \chi \mathrm{R} \psi$ and $\vdash \phi \mathrm{R} \chi \leftrightarrow \psi \mathrm{R} \chi$)

Although he does not specify explicitly, it is highly probable in the context of this and his prior work with Sosa [Chisholm and Sosa, 1966a; Chisholm and Sosa, 1966b] that he has classical truth-functional logic extended with propositional quantifiers in mind, and we will give this simple version:

(PL-1) All Tautologous REQ formulae
(PL-2) $\forall p_i \phi \rightarrow \phi(\psi/p)$, provided ψ is free for p in ϕ and ψ *is free of R*
(PL-3) $\forall p_i(\phi \rightarrow \psi) \rightarrow (\phi \rightarrow \forall p_i \psi)$, provided p is not free in ϕ
(MP) If $\vdash \phi$ and $\vdash \phi \rightarrow \psi$, then $\vdash \psi$
(UG) If $\vdash \phi$, $\vdash \forall p_i \phi$
(US) If $\vdash \phi$, then $\vdash \phi[\psi/\upsilon]$, where υ is a free variable in ϕ, and ψ is *free of R*

He also employs an alethic necessity operator, but he does not say what logic is to govern it. I will stipulate that we have the normal modal logic T for $\Box$:

(K) $\Box(\phi \rightarrow \psi) \rightarrow (\Box \phi \rightarrow \Box \psi)$
(T) $\Box \phi \rightarrow \phi$
(N) If $\vdash \phi$, then $\vdash \Box \phi$

He also speaks of performances and so I have included an operator for agency, $\mathbf{B}\mathrm{A}$, reminiscent of that referenced in [Chisholm, 1964]. More on this shortly.

Chisholm motivates seven axioms governing R[56]:

(A1) $p\mathrm{R}q \rightarrow \exists p \exists q p \mathrm{R} q$ [Not needed]
(A2) $p\mathrm{R}q \rightarrow \Box(p\mathrm{R}q)$
(A3) $p\mathrm{R}q \rightarrow \Diamond(p \ \& \ q)$
(A4) $\exists p \exists q \exists r[\Diamond(p \ \& \ q) \ \& \ p\mathrm{R}r \ \& \ q\mathrm{R}\neg r)]$
(A5) $(p\mathrm{R}r \ \& \ q\mathrm{R}r) \rightarrow (p \vee q)Rr$
(A6) $(p\mathrm{R}q \ \& \ p\mathrm{R}r) \rightarrow pR(q \ \& \ r)$
(A7) $(p \vee q)\mathrm{R}r \rightarrow (p\mathrm{R}r \vee q\mathrm{R}r)$

[56]The universal closure is intended for all but A4, and likewise for formulae with free propositional variables (p, q, r, s, and t).

Chisholm indicates that A1 is meant to reflect the commitment to propositions (or states of affairs). A1 is a logical truth in the intended system , so we must view A1 as emphatic, but redundant. A2 is meant to reflect the idea that the relation he seeks to capture is one that holds necessarily between its relata. A3 asserts that p would require q only if p and q are compatible, and he takes this to be "reflected in the principle that 'ought' implies 'can"' (p.5). A4 asserts that there are propositions p and q such that although p is compatible with q, each requires something directly contradicting what the other requires (and so conflicting requirements are possible). A5-A7 are interesting since they propose principles governing an operator's behavior (R's here) in the truth functional contexts of disjunction and conjunction, something not previously touched on in [Chisholm, 1964], nor in [Chisholm, 1963b], [Chisholm and Sosa, 1966b; Chisholm and Sosa, 1966a]. A5 states that if each of p and q would individually require r, then their disjunction would as well; A6 states that if p would individually require q and individually require r, then p would require their conjunction; and A7 states that if the disjunction of p with q would require r, then either p individually would require r or q would.

Given the presence of $\Box$ and its intended interpretation, we will strengthen the framework by endorsing an additional axiom, from which the weaker RE rule is derivable:

$$\text{(A8)} \quad \Box(p \leftrightarrow q) \rightarrow [(p\mathrm{R}s \leftrightarrow q\mathrm{R}s) \;\&\; (s\mathrm{R}p \leftrightarrow s\mathrm{R}q)]$$

In the process of expositing the key concepts and principles, he indicates that certain principles should *not* hold:

$$\begin{array}{ll} \nvdash & p\mathrm{R}q \rightarrow (p \vee q) \\ \nvdash & (s \;\&\; p\mathrm{R}q) \rightarrow (s \;\&\; p)\mathrm{R}q \\ \nvdash & (s \;\&\; p\mathrm{R}q) \rightarrow p\mathrm{R}(q \;\&\; s) \\ \nvdash & (p \vee q)\mathrm{R}r \rightarrow (p\mathrm{R}r \;\&\; q\mathrm{R}r) \end{array}$$

The first was noted in his 1964 account, but the second and third are new, and they assert, respectively, that neither the antecedent nor the consequent of a requirement relation can be (automatically) strengthened by something true. He argues that the fourth, which is the converse of A5, is too strong, and he suggests that A7 is a suitably weakened form (pp. 6-7).

Chisolm lists (exactly) five theorems in the article (we will identify many others):

(T1) $p\mathrm{R}q \rightarrow \Diamond p$
(T2) $p\mathrm{R}q \rightarrow \Diamond q$
(T3) $p\mathrm{R}q \rightarrow \neg(p\mathrm{R}\neg q)$ (from A3, A6)
(T4) $(p\mathrm{R}q \ \& \ r\mathrm{R}\neg q) \rightarrow (\neg[(p \ \& \ r)\mathrm{R}q] \vee \neg[(p \ \& \ r)\mathrm{R}\neg q]\ )$
(T5) $p\mathrm{R}q \rightarrow [(p \ \& \ r)\mathrm{R}q \vee (p \ \& \ \neg r)\mathrm{R}q]$

T1 & T2 indicate that the relata of the requirement relation must be individually possible; T3 rules out conflicting requirements generated by the same proposition; T4 states that if p and r generate conflicting requirements, then together, they will fail to generate at least one of the conflicting requirements (at least one will be defeated);[57] and T5 says that if p would require q, then either p conjoined with r would require q or p conjoined with $\neg r$ would require q. The following also hold, indicating what is logically impossible neither requires anything norWhat d is required by anything:

(T6) $\neg(\bot\mathrm{R}q)$
(T7) $\neg(q\mathrm{R}\bot)$

He notes that T4 implies that if p and r generate conflicting requirements, then it might be said that at least one is "overridden" (defeated) by their conjunction. Consistent with what Chisholm intends, we might represent this notion via a triadic operator for the subjunctive analog:

(D2) $\chi\mathbf{O}\mathrm{V}\phi\psi \stackrel{\text{def}}{=} \phi\mathrm{R}\psi \ \& \ \neg[(\phi \ \& \ \chi)\mathrm{R}\psi] \ \& \ \Diamond(\phi \ \& \ \psi \ \& \ \chi)$ [χ would override ϕ's requiring ψ]

Chisholm notes that the third conjunct is needed, roughly because otherwise the negation of what ϕ requires would always override what ϕ requires ($\neg\psi$ would override $\phi\mathrm{R}\psi$). We can encode the general point here as a theorem, along with a special case:

(T8) $\neg[(p \ \& \ \neg q)\mathrm{R}q]$
(T9) $\neg(p\mathrm{R}\neg p \vee \neg p\mathrm{R}p)$

With this, Chisholm turns to laying out an analytic scheme much like the one in 1964. The de facto version of D2 is:

[57] The antecedent is redundant, so we must view it as merely emphatic, stressing that conflicting requirements resolve.

(D3) $\chi\mathbf{O}\mathrm{V}'\phi\psi \stackrel{\text{def}}{=} \phi\mathrm{R}\psi \;\&\; \neg[(\phi \;\&\; \chi)\mathrm{R}\psi] \;\&\; \phi \;\&\; \chi$ [χ de facto overrides ϕ's requiring ψ]

Chisholm here notes again that an "overriding may itself be overridden." Chisholm then reintroduces an analysis of what ought to be (all things considered):

(D4) $\mathbf{O}\mathrm{U}^{\mathrm{b}}\phi \stackrel{\text{def}}{=} \exists\mathrm{p}(\mathrm{pR}'\phi \;\&\; \neg\exists\mathrm{r}[\mathrm{r} \;\&\; (\mathrm{r} \;\&\; \mathrm{p})\mathbf{O}\mathrm{V}'\mathrm{p}\phi])$

What does this amount to in primitive notation? Essentially this:

(T10) $\mathbf{O}\mathrm{U}^{\mathrm{b}}\phi \leftrightarrow \exists p[p \;\&\; p\mathrm{R}\phi \;\&\; \forall r(r \rightarrow (p \;\&\; r)\mathrm{R}\phi)]$

So $\mathbf{O}\mathrm{U}^{\mathrm{b}}$ represents what overridingly ought to be. It ought to be that ϕ iff something de facto requires ϕ and anything true conjoined with it requires ϕ. As in the axiological frameworks for $\mathbf{O}$B [Chisholm, 1963b; Chisholm and Sosa, 1966b], Chisholm clearly intends that $\mathbf{O}\mathrm{U}^{\mathrm{b}}$ not be subject to conflicts, and this is derivable as well:

(T11) $\mathbf{O}\mathrm{U}^{\mathrm{b}}q \rightarrow \neg\mathbf{O}\mathrm{U}^{\mathrm{b}}\neg q$

This can be generalized to include non-explicit conflicts described via the possibility operator:

(T12) $\neg\Diamond(p \;\&\; q) \rightarrow \neg(\mathbf{O}\mathrm{U}^{\mathrm{b}}p \;\&\; \mathbf{O}\mathrm{U}^{\mathrm{b}}q)$

Similarly, an RE principle is derivable for this operator as well:

(RE-$\mathbf{O}\mathrm{U}^{\mathrm{b}}$) If $\vdash \phi \leftrightarrow \psi$ then $\vdash \mathbf{O}\mathrm{U}^{\mathrm{b}}\phi \leftrightarrow \mathbf{O}\mathrm{U}^{\mathrm{b}}\psi$

Chisholm next uses his requirement framework to analyze *prima facie duty*, in the process shifting to talk of agency, although he does not explicitly introduce an agency operator as he did in 1964, but instead informally uses the language of actions mixed in with the above operators. We will regiment here by using the agency operator as in the 1964 presentation, making things a bit more precise about its logic in a moment:

(D5) $\mathbf{P}\mathrm{F}\phi \stackrel{\text{def}}{=} \exists q(q\mathrm{R}'\mathbf{B}\mathrm{A}\phi)$. [S has a *prima facie* duty to act so that ϕ]

(D6) $\mathbf{P}\mathrm{F}'^{*}\phi \stackrel{\text{def}}{=} \exists q \exists r(q\mathrm{R}'\mathbf{B}\mathrm{A}r \;\&\; \neg\Diamond(\mathbf{B}\mathrm{A}r \;\&\; \mathbf{B}\mathrm{A}\phi))$ [Unofficial def: S has a *prima facie* prohibition against bringing about ϕ]

So Jane Doe has a prima facie obligation to bring it about that ϕ iff something true requires that she do so; and Jane Doe has a prima facie prohibition against bringing it about that ϕ iff something true requires her to bring about some proposition, the bringing about of which is incompatible with her bringing about ϕ. (The "*" in D6 indicates a difficulty, which we will specify momentarily.) One might wonder why Chisholm does not try to subsume the notion of prima facie prohibition against bringing about ϕ to a requirement to not do or omit ϕ: $\mathbf{P}\mathrm{F}'^{*}\phi$ iff $\exists q(q\mathrm{R}'\neg\mathbf{B}\mathrm{A}\phi)$? Perhaps he thinks this does not require anything agential of Jane.[58] But then one wonders why does Chisholm not just subsume prima facie *prohibition* against bringing about ϕ to a special case of a positive prima facie *duty* to bring it about that you don't bring about ϕ (i.e. to refrain from ϕ):

$\mathbf{P}\mathrm{F}'^{*}\phi$ as $\exists q(q\mathrm{R}'\mathbf{B}\mathrm{A}\neg\mathbf{B}\mathrm{A}\phi)$?

Leaving these questions aside, the problem with D6 is reflected in this derived rule:

(R$\mathbf{P}\mathrm{F}'^{*}$) If $\vdash \neg\phi$, then $\vdash \exists q \exists r(q\mathrm{R}'\mathbf{B}\mathrm{A}r) \rightarrow \mathbf{P}\mathrm{F}'^{*}\mathbf{B}\mathrm{A}\phi)$

That is, if ϕ is a logical falsehood (e.g. $\bot$) and there is a de facto requirement for Jane to bring about anything at all, then there is a prima facie prohibition for Jane against bringing about that logical falsehood per D6. There lies the flaw: one can't bring it about that one does not bring it about that $\bot$ (and so one can't do that *by* doing some other thing, r), since

(T13) $\neg\mathbf{B}\mathrm{A}\neg\mathbf{B}\mathrm{A}\bot$

is plausibly deemed a logical truth (and labelled a theorem in anticipation), and so $\neg\mathbf{B}\mathrm{A}\bot$ is not something I can bring about by bringing

[58] This issue will resurface later on when we consider supererogatory and offensive omissions.

about something else.

Since R$\mathbf{P}$F$'^{*}$ is an odd consequence and probably not desired by Chisholm, we will add one more conjunct to D6's definiens assuring that it is logically possible to bring about anything prima facie prohibited:

(D7) $\mathbf{P}\mathrm{F}'\phi \stackrel{\text{def}}{=} \exists q \exists r(q\mathrm{R}'\mathbf{B}\mathrm{A}r \;\&\; \neg\Diamond(\mathbf{B}\mathrm{A}r \;\&\; \mathbf{B}\mathrm{A}\phi)) \;\&\; \Diamond\mathbf{B}\mathrm{A}\phi$. [S has a *prima facie* prohibition against bringing it about that ϕ]

Let this be the official definition of the notion Chisholm intends.

Chisholm does not specify what the logic for agency might look like, beyond his mention of Anselm and the square of opposition in his 1964 rendition of the requirement framework, which is too thin.[59] We will need to rely on some basic logical properties for $\mathbf{B}$A. So let's stipulate that the following minor modification (to account for presence of a necessity operator) of a familiar simple system is incorporated:

(T) $\vdash \mathbf{B}\mathrm{A}p \rightarrow p$
(C) $\vdash (\mathbf{B}\mathrm{A}p \;\&\; \mathbf{B}\mathrm{A}q) \rightarrow \mathbf{B}\mathrm{A}(p \;\&\; q)$
(NO-$\Box$) $\vdash \Box p \rightarrow \neg\mathbf{B}\mathrm{A}p$
(REN) $\vdash \Box(p \leftrightarrow q) \rightarrow (\mathbf{B}\mathrm{A}p \leftrightarrow \mathbf{B}\mathrm{A}q)$

The last two items link $\mathbf{B}$A to $\Box$, and then easily allow us to derive a thesis and rule that are often presented with T and C as constituting a core logic for agency when no necessity operator is present:

(T14) $\neg\mathbf{B}\mathrm{A}\top$ [NO]
(RE-BA) If $\vdash \phi \leftrightarrow \psi$ then $\vdash \mathbf{B}\mathrm{A}\phi \leftrightarrow \mathbf{B}\mathrm{A}\psi$[60]

We will make free use of these in developing and elaborating Chisholm's framework below, and we hope the reader will forgive our making slight anticipatory use of this system above.

One immediate question raised is *how are R and* $\boldsymbol{BA}$ *to be linked*? Chisholm does not raise this issue at all, which is unfortunate, since

[59] Recall EQ from Section 1.2, where the square of opposition for $\mathbf{O}$B is tautologically equivalent to a no conflicts principle for $\mathbf{O}$B, and here we have just a notational variant, and although no conflicts for $\mathbf{B}$A is certainly a sound logical feature it is neither basic nor enough.

[60] See [Jones and Sergot, 1996] for a classical source. It should be noted that. the system above consisting of T, C, T17 (NO) and RE-BA (mnemonic: TECNO) ultimately derives from Elgesem's rich work on agency ([Elgesem, 1993], [Elgesem, 1997]). See also the discussion in [Governatori and Rotolo, 2005].

answering it matters, and as we've just seen, it already matters for his first introduction proper of reference to agency/action in the article we are expositing. Let's note first a theorem that reflects the intended interpretation of the requirement relation expressed in A2 as a relation holding between propositions solely based on the propositional content alone, so not something agents can impact:

(T15) $\neg\mathbf{B}\mathrm{A}(p\mathrm{R}q)$

Note that this is rather important for restricting the conception of "would require"/"would call for" here. For there are certainly some senses in which an agent can bring it about that some p would, or even does, require some proposition q. Like a number of ethicists, Chisholm thinks some links hold necessarily, like promise-making requires promise-keeping, the former *intrinsically requiring* the latter. Regarding agency and requirement, these also follow readily:

(T16) $p\mathrm{R}\mathbf{B}\mathrm{A}q \rightarrow \neg(p\mathrm{R}\neg\mathbf{B}\mathrm{A}q)$
(T17) $p\mathrm{R}\mathbf{B}\mathrm{A}q \rightarrow \neg(p\mathrm{R}\mathbf{B}\mathrm{A}\neg q)$
(T18) $p\mathrm{R}\mathbf{B}\mathrm{A}q \rightarrow \neg(p\mathrm{R}\neg q)$

However, although it is crucial to Chisholm's framework that something non-agential can be such that it ought to be, and thus is impersonally strictly required, without it being the case that Jane Doe ought to bring that thing about, it is very plausible to think that what an agent overridingly ought to make true (which by definition is determined at the impersonal level) is itself impersonally required by the situation. We will add an axiom that allows us to generate a proof of the just mentioned ought to do principle: namely that if p would require that I bring something about, then p would also require what it requires me to bring about:

(B1) $p\mathrm{R}\mathbf{B}\mathrm{A}q \rightarrow p\mathrm{R}q$.[61]

Here is another potential axiom,

(B2) $p\mathrm{R}\neg q \rightarrow p\mathrm{R}\neg\mathbf{B}\mathrm{A}q$ [equivalently, $p\mathrm{R}q \rightarrow p\mathrm{R}\neg\mathbf{B}\mathrm{A}\neg p$]

[61]Note that this does not have the consequence that whatever follows from what I am required to bring about is required (which would result in $p\mathrm{R}q \rightarrow p\mathrm{R}\top$), for B1 is confined to what the agent can bring about (and $\neg\mathbf{B}\mathrm{A}\top$ is a thesis).

saying that if p requires q's absence, then it requires as well the absence of my making q present. B2 serves in supporting the following:

$$(\text{T19}^{\text{B2}}) \quad \mathbf{O}\text{U}^{\text{b}}\neg p \rightarrow \mathbf{O}\text{U}^{\text{b}}\neg\mathbf{B}\text{A}p$$
$$[\text{equivalently, } \mathbf{O}\text{U}^{\text{b}}p \rightarrow \mathbf{O}\text{U}^{\text{b}}\neg\mathbf{B}\text{A}\neg p]$$

which has as an immediate corollary, this equivalent, where $\mathbf{P}\text{E}^{b}\phi \stackrel{\text{def}}{=} \neg\mathbf{O}\text{U}^{\text{b}}\neg\phi$:

$$(\text{T20}^{\text{B2}}) \quad \mathbf{P}\text{E}^{b}\mathbf{B}\text{A}p \rightarrow \mathbf{P}\text{E}^{b}p$$

B2 seems plausible. If p would require that q *not be true*, then p would require that Jane not make q true. Similar for T19^{B2}: if it overridingly ought to be that p is false, then it overridingly ought to be that I do not make p true.

Let me note that although B2 is not considered by Chisholm (no agency links are, alas), we will see later on that it has particularly untoward implications in the context of Chisholm's intended applications.

Another potential axiom that will be used near the end is

$$(\text{B3}) \quad p\text{R}\mathbf{B}\text{A}q \rightarrow p\text{R}\neg\mathbf{B}\text{A}\neg\mathbf{B}\text{A}q$$

stating that if p requires my bringing about q then it requires my not bringing it about that I do not bring about q.

Let me note one other possible axiom linking R to **B**A:

$$(\text{B4}) \quad p\text{R}\mathbf{B}\text{A}\neg q \rightarrow p\text{R}\neg\mathbf{B}\text{A}q$$

stating that if p requires that I bring about $\neg q$, then p requires that I not bring about q.[62] With B4 we can derive the following theorem, but apparently not without it:

$$(\text{T21}^{\text{B4}}) \quad \mathbf{O}\text{U}^{\text{b}}\mathbf{B}\text{A}p \rightarrow \mathbf{O}\text{U}^{\text{b}}\neg\mathbf{B}\text{A}\neg p$$

This theorem tells us that if it (overridingly) ought to be that Jane brings about p, then it ought to be that she does not bring about $\neg p$.

We will see that some of these potential axioms linking requirement to agency are of use in showing things he would likely countenance, oth-

[62]It might be thought that this is easy to prove, but it appears we can only get a not so close cousin: since $\vdash \mathbf{B}\text{A}\neg q \leftrightarrow (\mathbf{B}\text{A}\neg q \;\&\; \neg\mathbf{B}\text{A}q)$, given **B**A's T axiom, from $p\text{R}\mathbf{B}\text{A}\neg q$, it follows that $p\text{R}(\mathbf{B}\text{A}\neg q \;\&\; \neg\mathbf{B}\text{A}q)$; but this is not quite enough.

ers are either reductive or even disruptive. Thus I will consider the core system to not include any of B1-B4, and I will provide clear indications when a theorem depends on any of these additional potential axioms, for example as T21$^{\mathrm{B4}}$ just above does, which should then be read more strictly as a rule saying if B2 is added as a thesis to the core system, then so is the formulae to the right of the label. I hope this bit of sloppiness will be tolerable, since the conditionality on any of B1-B4 is clearly indicated.

Finally, Chisholm provides what is now often called the "Meinong-Chisholm Reduction", a proposed reduction of *agential oughts* to *impersonal oughts* plus agency:

(D8) $\mathbf{O}\mathrm{U}^{\mathrm{d}}\phi \stackrel{\mathrm{def}}{=} \mathbf{O}\mathrm{U}^{\mathrm{b}}\mathbf{B}\mathrm{A}\phi$ [Jane (overridingly) ought to act so that ϕ as it ought to be the case that she does]

Given D8, the following is now an immediate corollary of T11:

(T22) $\mathbf{O}\mathrm{U}^{\mathrm{d}}q \rightarrow \neg\mathbf{O}\mathrm{U}^{\mathrm{b}}\neg\mathbf{B}\mathrm{A}q$

Although not noted, but likely welcomed, replacement by provable equivalents follows for $\mathbf{O}\mathrm{U}^{\mathrm{d}}$:

(RE-$\mathbf{O}\mathrm{U}^{\mathrm{d}}$) If $\vdash \phi \leftrightarrow \psi$ then $\vdash \mathbf{O}\mathrm{U}^{\mathrm{d}}\phi \leftrightarrow \mathbf{O}\mathrm{U}^{\mathrm{d}}\psi$

Furthermore, as suggested above when introducing B1, it is plausible to think that what I ought to make true, ought to be true, and this is provable, *given* B1:

(T23$^{\mathrm{B1}}$) $\mathbf{O}\mathrm{U}^{\mathrm{d}}q \rightarrow \mathbf{O}\mathrm{U}^{\mathrm{b}}q$

which has this trivial corollary,

(T24$^{\mathrm{B1}}$) $\mathbf{P}\mathrm{E}^{\mathrm{b}}q \rightarrow \neg\mathbf{O}\mathrm{U}^{\mathrm{d}}\neg q$

Let us note here that versions of *ought implies can* follow:

(T25) $\mathbf{O}\mathrm{U}^{\mathrm{b}}p \rightarrow \Diamond p$
(T26) $\mathbf{O}\mathrm{U}^{\mathrm{d}}p \rightarrow (\Diamond\mathbf{B}\mathrm{A}p \,\&\, \Diamond p)$
(T27) $\mathbf{P}\mathrm{F}p \rightarrow (\Diamond\mathbf{B}\mathrm{A}p \,\&\, \Diamond p)$
(T28) $\mathbf{P}\mathrm{F}'p \rightarrow (\Diamond\mathbf{B}\mathrm{A}p \,\&\, \Diamond\neg\mathbf{B}\mathrm{A}p \,\&\, \Diamond p)$

With this exposition of the core of the 1964 and 1974 requirement framework, we turn to his handling of supererogation and kin, ignoring the other seven problems he addressed in 1964, most of which are addressed as well in 1974.

4.3 Application and extension of the framework to supererogation and kin

As mentioned above, in Chisholm's 1964 articulation [Chisholm, 1964] of the conceptual framework for requirement, he explicitly discussed the application of his analytic framework to supererogation and kin, and does *not* do so in the [Chisholm, 1974] article itself, but only in responding to two commentators on the article, and thus less systematically. I will begin with the 1964 discussion.

The 1964 explicit application

To prepare the way for applying the central framework to supererogation and kin in [Chisholm, 1964], he begins by defining a dual for *ought to be* (as we did above in passing), which he glosses as "permitted", saying that this is more stipulation than analysis:

(D9) $\mathbf{P}\mathrm{E}^{\mathrm{b}}\phi \stackrel{\mathrm{def}}{=} \neg\mathbf{O}\mathrm{U}^{\mathrm{b}}\neg\phi$
[It is "impersonally permissible" that ϕ]

An obvious corollary of T11 given D9 is:

(T29) $\mathbf{O}\mathrm{U}^{\mathrm{b}}p \rightarrow \mathbf{P}\mathrm{E}^{\mathrm{b}}p$

The following equivalence is also easily derivable:

(T30) $\mathbf{P}\mathrm{E}^{\mathrm{b}}q \leftrightarrow \forall p[p \;\&\; p\mathrm{R}\neg q. \rightarrow \exists r(r \;\&\; \neg((p \;\&\; r)\mathrm{R}\neg q))]$

stating essentially that q is impersonally permissible iff any p that de facto requires q's absence is defeated by some true expansion of p — so q is ultimately clear of any sustainable exclusion. A replacement rule for $\mathbf{P}\mathrm{E}^{\mathrm{b}}$ also readily follows:

(RE-$\mathbf{P}\mathrm{E}^{\mathrm{b}}$) If $\vdash \phi \leftrightarrow \psi$, then $\vdash \mathbf{P}\mathrm{E}^{\mathrm{b}}\phi \leftrightarrow \mathbf{P}\mathrm{E}^{\mathrm{b}}\psi$

Chisholm states that permission proper, permission to do, amounts to the absence of a recommendation against it – it's not being the case that it ought to not be done, which we will represent agentially as follows:

(D10) $\mathbf{P}\mathrm{E}^{\mathrm{d}}\phi \stackrel{\mathrm{def}}{=} \mathbf{P}\mathrm{E}^{\mathrm{b}}\mathbf{B}\mathrm{A}\phi$. [It is permissible for Jane to act ("do") so that ϕ]

So ϕ is agentially permissible for S iff it is impersonally permissible that S brings it about that ϕ.

Following his informal gloss from [Chisholm, 1963b], that the supererogatory is "non-obligatory well-doing"[63], he provides the following requirement framework definition:

(D11) $\mathbf{S}\mathrm{U}\mathbf{B}\mathrm{A}\phi \stackrel{\mathrm{def}}{=} \mathbf{O}\mathrm{U}^{\mathrm{b}}\phi \,\&\, \mathbf{P}\mathrm{E}^{\mathrm{b}}\neg\mathbf{B}\mathrm{A}\phi \,\&\, \mathbf{P}\mathrm{E}^{\mathrm{b}}\mathbf{B}\mathrm{A}\phi$. [It is supererogatory to act so that ϕ][64]

So it is supererogatory for S to bring it about that ϕ iff ϕ itself ought to be the case, but it is it is not the case that it ought to be that S brings about ϕ nor is it the case that it ought to be that S does not bring about ϕ. So the idea is that something is supererogatory if it ought to obtain, but there is no onus on S to make it so nor to not make it so. A classic example is that of one volunteering in a group to go on a dangerous mission. Chisholm leaves open the question of whether or not the last conjunct of the definiens is redundant, presumably because he thinks it is perhaps necessitated by the first conjunct. However, it is indeed needed. For there will be many cases where something ought to be the case and it also ought not involve my agency at all–it is out of my jurisdiction (e.g. disciplining your child), so without the third conjunct, all such misplaced exercises of agency would be supererogatory.

Chisholm once again glosses an offence as a case of "permissive ill-doing", and he defines it as the symmetrical analog of supererogation:

(D12) $\mathbf{O}\mathrm{F}\mathbf{B}\mathrm{A}\phi \stackrel{\mathrm{def}}{=} \mathbf{O}\mathrm{U}^{\mathrm{b}}\neg\phi \,\&\, \mathbf{P}\mathrm{E}^{\mathrm{b}}\mathbf{B}\mathrm{A}\phi \,\&\, \mathbf{P}\mathrm{E}^{\mathrm{b}}\neg\mathbf{B}\mathrm{A}\phi$. [It is an offence for S to act so that ϕ]

So it is an offence for S to bring about ϕ iff $\neg\phi$ ought to be the case, but it is impersonally permissible that S bring about ϕ, and it is impersonally

[63] In his 1964 rendition, the first clause of the definiens of D11 is identified with *it is good that* ϕ, and Chisholm endorses this in 1974 as well, as we will see.

[64] Reminder as elsewhere above and below, "is" might be better read as "would be".

permissible that S does not.

Chisholm rounds things out by providing an analysis of *optionality* and *indifference* that does not conflate the two (as has so often been done in deontic logic and ethical theory). He characterizes *agential optionality*, the optionality of S's bringing it about that ϕ, as follows:

(D13) $\mathbf{O}\mathrm{P}^{\mathrm{d}}\phi \stackrel{\mathrm{def}}{=} \mathbf{P}\mathrm{E}^{\mathrm{b}}\mathbf{B}\mathrm{A}\phi \;\&\; \mathbf{P}\mathrm{E}^{\mathrm{b}}\neg\mathbf{B}\mathrm{A}\phi$. [It is agentially optional that S bring it about that ϕ]

Although Chisholm does not articulate this, he could define "impersonal optionality", in a natural manner consistent with his characterization of impersonal permissibility, as follows:

(D14) $\mathbf{O}\mathrm{P}^{\mathrm{b}}\phi \stackrel{\mathrm{def}}{=} \mathbf{P}\mathrm{E}^{\mathrm{b}}\phi \;\&\; \mathbf{P}\mathrm{E}^{\mathrm{b}}\neg\phi$ [It is impersonally optional that ϕ]

Then optionality of one's agency regarding ϕ would be a special case,

(T31) $\mathbf{O}\mathrm{P}^{\mathrm{d}}p \leftrightarrow \mathbf{O}\mathrm{P}^{\mathrm{b}}\mathbf{B}\mathrm{A}p$

It is also the case that the indifference of impersonal optionality to negation follows:

(T32) $\mathbf{O}\mathrm{P}^{\mathrm{b}}\phi \leftrightarrow \mathbf{O}\mathrm{P}^{\mathrm{b}}\neg p$

Although this indifference to negation does not hold for $\mathbf{O}\mathrm{P}^{\mathrm{d}}$, and should not, since it might be that I can bring about p or not do so, but it might be wrong for me to bring about $\neg p$ — to essentially prevent p. The following desirable theorem does hold, and is just a special case of the preceding theorem:

(T33) $\mathbf{O}\mathrm{P}^{\mathrm{b}}\mathbf{B}\mathrm{A}p \leftrightarrow \mathbf{O}\mathrm{P}^{\mathrm{b}}\neg\mathbf{B}\mathrm{A}p$

Chisholm then characterizes *agential indifference* regarding ϕ as follows:

(D15) $\mathbf{I}\mathrm{N}\mathbf{B}\mathrm{A}\phi \stackrel{\mathrm{def}}{=} \mathbf{P}\mathrm{E}^{\mathrm{b}}\phi \;\&\; \mathbf{P}\mathrm{E}^{\mathrm{b}}\neg\phi \;\&\; \mathbf{P}\mathrm{E}^{\mathrm{b}}\mathbf{B}\mathrm{A}\phi \;\&\; \mathbf{P}\mathrm{E}^{\mathrm{b}}\neg\mathbf{B}\mathrm{A}\phi$. [It is indifferent that S act so that ϕ.]

That is, it is a matter of indifference for our agent to bring about ϕ iff it is both impersonally optional that ϕ and agentially optional that ϕ:

(T34) $\mathbf{IN}\mathbf{B}\mathrm{A}p \leftrightarrow (\mathbf{O}\mathrm{P}^{\mathrm{b}}p \;\&\; \mathbf{O}\mathrm{P}^{\mathrm{d}}p)$

and thus another desirable result follows: indifference *properly* implies optionality (an essential result in accommodating supererogation or offences), and in this framework, this is so in both senses of optionality:

(T35) $(\mathbf{IN}\mathbf{B}\mathrm{A}p \rightarrow \mathbf{O}\mathrm{P}^{\mathrm{b}}p) \;\&\; (\mathbf{IN}\mathbf{B}\mathrm{A}p \rightarrow \mathbf{O}\mathrm{P}^{\mathrm{d}}p)$[65]

However, note that we do not have an *impersonal* analog to agential indifference as we do have for agential optionality. At the impersonal level, there is no distinction between indifference and optionality in Chisholm's REQ framework (unlike, for example, in the axiology-based framework of Chisholm-Sosa).

We argued that indifference, like optionality, should be indifferent to negation. Is it as represented here? Again, as with optionality, the answer is no, and as defined, it should be 'no'. For we have no independent characterization of $\mathbf{IN}$ per se, only the indifference of bringing something about, $\mathbf{IN}\mathbf{B}\mathrm{A}\phi$ as an abbreviation for ($\mathbf{P}\mathrm{E}^{\mathrm{b}}\phi$ & $\mathbf{P}\mathrm{E}^{\mathrm{b}}\neg\phi$ & $\mathbf{P}\mathrm{E}^{\mathrm{b}}64\mathbf{B}\mathrm{A}\phi$ & $\mathbf{P}\mathrm{E}^{\mathrm{b}}\neg\mathbf{B}\mathrm{A}\phi$), and thus we have no representation of the indifference of not bringing something about, but the conditions ought to be essentially the same, and so we will stipulate:

(D16) $\mathbf{IN}\neg\mathbf{B}\mathrm{A}\phi \stackrel{\text{def}}{=} (\mathbf{P}\mathrm{E}^{\mathrm{b}}\neg\phi \;\&\; \mathbf{P}\mathrm{E}^{\mathrm{b}}\phi \;\&\; \mathbf{P}\mathrm{E}^{\mathrm{b}}\neg\mathbf{B}\mathrm{A}\phi \;\&\; \mathbf{P}\mathrm{E}^{\mathrm{b}}\mathbf{B}\mathrm{A}\phi)$

The definiens of D16 is essentially just the definiens of D15 with the conjuncts in a different order for emphasis, and so the indifference of *agential* indifference to negation follows trivially, essentially by stipulation:

(T36) $\mathbf{IN}\mathbf{B}\mathrm{A}p \leftrightarrow \mathbf{IN}\neg\mathbf{B}\mathrm{A}p$

An indifference exclusion principle regarding what ought to be follows:

(T37) $\mathbf{IN}\mathbf{B}\mathrm{A}p \rightarrow (\neg\mathbf{O}\mathrm{U}^{\mathrm{b}}p \;\&\; \neg\mathbf{O}\mathrm{U}^{\mathrm{b}}\neg p \;\&\; \neg\mathbf{O}\mathrm{U}^{\mathrm{b}}\mathbf{B}\mathrm{A}p \;\&\; \neg\mathbf{O}\mathrm{U}\neg\mathbf{B}\mathrm{A}p)$

With the optionality notions defined, we can then note a more concise way of expressing the intended accounts of supererogation and offence:

[65] We pass over a semantics where falsifying models would show the implications are proper.

(T38) $\mathbf{SUBA}p \leftrightarrow (\mathbf{O}\mathrm{U}^{\mathrm{b}}p \;\&\; \mathbf{O}\mathrm{P}^{\mathrm{d}}p)$
(T39) $\mathbf{OFBA}p \leftrightarrow (\mathbf{O}\mathrm{U}^{\mathrm{b}}\neg\mathrm{p} \;\&\; \mathbf{O}\mathrm{P}^{\mathrm{d}}p)$

The 1974 discussion of applications

In [Chisholm, 1974] his article mentions essentially the same problems that he hopes his requirement framework can solve as those mentioned in [Chisholm, 1964], except for the conspicuous absence of the problem of supererogation and kin, which is the problem discussed at greatest length in the 1964 piece. However, the volume that [Chisholm, 1974] appears in contains commentary on his article and his replies. In this commentary, Anscombe and Raz both suggest that the account is not consistent with the possibility of supererogation. Here, it appears that neither critic is aware of the 1964 account. Chisholm responds that, to the contrary, he believes his requirement framework can account for supererogation and he cites his "non-obligatory well-doing" gloss that we've seen before in [Chisholm, 1963b], [Chisholm, 1964], and [Chisholm and Sosa, 1966a], and he goes on to unpack that gloss in the same way he did in [Chisholm, 1964], regimented in our D11 above, *explicitly using the bring it about that agency idiom* (rather than doing/not doing):

$$\mathbf{SUBA}\phi \stackrel{\text{def}}{=} \mathbf{O}\mathrm{U}^{\mathrm{b}}\phi \;\&\; \mathbf{P}\mathrm{E}^{\mathrm{b}}\neg\mathbf{BA}\phi \;\&\; \mathbf{P}\mathrm{E}^{\mathrm{b}}\mathbf{BA}\phi. \text{ [D11 above]}$$

There is some additional discussion in the commentary, but it does not invoke any clarification of the formal account of supererogation, just a defense against an objection.

Compared to the explicit exposition in [Chisholm, 1964], this is thin coverage, and it only comes when prompted in the commentary. Still, I think what is said there indicates that Chisholm holds the same basic view about the requirement framework's application to supererogation and kindred notions as in [Chisholm, 1964].

4.4 Comparison with the prior frameworks and some challenges/disruptions.

His account of good and bad in his 1964 and 1974 articulation of the REQ framework match:

(D17) $\mathbf{GD}\phi \stackrel{\text{def}}{=} \mathbf{O}\mathrm{U}^{\mathrm{b}}\phi$ [It is, or would be, good that ϕ]
(D18) $\mathbf{BD}\phi \stackrel{\text{def}}{=} \mathbf{O}\mathrm{U}^{\mathrm{b}}\neg\phi$ [It is, or would be, bad that ϕ]

RE principles follow for each from those for $\mathbf{O}\mathrm{U}^{\mathrm{b}}$:

(RE-$\mathbf{G}$D/$\mathbf{B}$D) If $\vdash \phi \leftrightarrow \psi$ then $\vdash \mathbf{G}\mathrm{D}\phi \leftrightarrow \mathbf{G}\mathrm{D}\psi$ and $\vdash \mathbf{B}\mathrm{D}\phi \leftrightarrow \mathbf{B}\mathrm{D}\psi$

Chisholm's main thesis, CCT, for his good-bad axiological framework follows:

(T40) $\neg(\mathbf{G}\mathrm{D}p \,\&\, \mathbf{B}\mathrm{D}p)$ (Chisholm's Contrariety Thesis)

However, it is also clear that nothing can be totally supererogatory or totally offensive, that is, the defining conditions for those concepts in [Chisholm, 1963b] are ruled out as logically impossible given the analysis of *good* and *bad*, and *ought to be* given here:

(T41) $\neg(\mathbf{G}\mathrm{D}p \,\&\, \mathbf{G}\mathrm{D}\neg p)$ [NC-GD]
(T42) $\neg(\mathbf{B}\mathrm{D}p \,\&\, \mathbf{B}\mathrm{D}\neg p)$ [NC-BD]
(T43) $\neg(\mathbf{T}\mathrm{S} \vee \mathbf{T}\mathrm{O}p)$ [Totally supererogatory/offensive]

Thus we have a logic for GD and BD at least as strong as RCGB[6]. T43 is perhaps not surprising, for as mentioned earlier, the Chisholm-Sosa framework of 1964, also entailed these no-conflict principles for $\mathbf{G}$D and for $\mathbf{B}$D individually, and so ruled out the totally supererogatory and the totally offensive. Only the [Chisholm, 1963b] scheme made a place for them.

More significantly, one of the key constraints from [Chisholm, 1963b], Chisholm Independence Constraint

(CIC) $\nvdash \mathbf{G}\mathrm{D}p \rightarrow \mathbf{B}\mathrm{D}\neg p$,
$\nvdash \mathbf{B}\mathrm{D}\neg p \rightarrow \mathbf{G}\mathrm{D}p$,
$\nvdash \mathbf{G}\mathrm{D}\neg p \rightarrow \mathbf{B}\mathrm{D}p$, and
$\nvdash \mathbf{B}\mathrm{D}p \rightarrow \mathbf{G}\mathrm{D}\neg p$

is not sustainable here, for each formulae is obviously a theorem in the REQ framework:

(T44) $\mathbf{G}\mathrm{D}p \leftrightarrow \mathbf{B}\mathrm{D}\neg p$,
(T45) $\mathbf{G}\mathrm{D}\neg p \leftrightarrow \mathbf{B}\mathrm{D}p$

and so we wind up with a logical framework for good and bad that is stronger than Chisholm explicitly provided for in his 1963 framework, and T44 and T45 are also not theorems of the Chisholm-Sosa framework. Hence we begin to diverge from Chisholm's 1963 framework consider-

ably here. Recall that in that scheme, the following analyses are offered (where, recall, neutrality is defined so that $\mathbf{N}\mathrm{U}\phi \leftrightarrow (\neg\mathbf{G}\mathrm{D}\phi \;\&\; \neg\mathbf{B}\mathrm{D}\phi)$:

(CDS′)	
Totally Offensive (**T**O):	$\mathbf{T}\mathrm{O}\phi \stackrel{\text{def}}{=} \mathbf{B}\mathrm{D}\phi \;\&\; \mathbf{B}\mathrm{D}\neg\phi$
Offence of Commission (**O**F)	$\mathbf{O}\mathrm{F}\phi \stackrel{\text{def}}{=} \mathbf{B}\mathrm{D}\phi \;\&\; \mathbf{N}\mathrm{U}\neg\phi$
Forbidden (**I**M)	$\mathbf{I}\mathrm{M}\phi \stackrel{\text{def}}{=} \mathbf{B}\mathrm{D}\phi \;\&\; \mathbf{G}\mathrm{D}\neg\phi$
Offence of Omission (**O**F¬)	$\mathbf{O}\mathrm{F}\neg\phi \stackrel{\text{def}}{=} \mathbf{B}\mathrm{D}\neg\phi \;\&\; \mathbf{N}\mathrm{U}\phi$
Indifferent (**I**N)	$\mathbf{I}\mathrm{N}\phi \stackrel{\text{def}}{=} \mathbf{N}\mathrm{U}\phi \;\&\; \mathbf{N}\mathrm{U}\neg\phi$
Supererogatory Omission (**S**U¬)	$\mathbf{S}\mathrm{U}\neg\phi \stackrel{\text{def}}{=} \mathbf{G}\mathrm{D}\neg\phi \;\&\; \mathbf{N}\mathrm{U}\phi$
Obligatory (**O**B)	$\mathbf{O}\mathrm{B}\phi \stackrel{\text{def}}{=} \mathbf{G}\mathrm{D}\phi \;\&\; \mathbf{B}\mathrm{D}\neg\phi$
Supererogatory Commission (**S**U)	$\mathbf{S}\mathrm{U}\phi \stackrel{\text{def}}{=} \mathbf{G}\mathrm{D}\phi \;\&\; \mathbf{N}\mathrm{U}\neg\phi$
Totally Supererogatory (**T**S)	$\mathbf{T}\mathrm{S}\phi \stackrel{\text{def}}{=} \mathbf{G}\mathrm{D}\phi \;\&\; \mathbf{G}\mathrm{D}\neg\phi$

Not only are the first and last ruled out (T41 and T42), but as we noticed in expositing "Supererogation and Offence", various redundancies as well as incoherent combinations would emerge if the equivalences implied above by the definitions were maintained. Consider the trivially equivalent formulations below, where strike-throughs express redundancies that could be deleted without loss, and underlining indicates now-inconsistent conditions:

Supererogatory Commission (**S**U)	$\underline{\mathbf{S}\mathrm{U}\phi} \stackrel{\text{def}}{=} \underline{\mathbf{G}\mathrm{D}\phi \;\&\; \neg\mathbf{B}\mathrm{D}\neg\phi \;\&\; \neg\mathbf{G}\mathrm{D}\neg\phi}$
Obligatory (**O**B)	$\mathbf{O}\mathrm{B}\phi \stackrel{\text{def}}{=} \mathbf{G}\mathrm{D}\phi \;\cancel{\&\; \mathbf{B}\mathrm{D}\neg\phi}$
Supererogatory Omission (SU¬)	$\underline{\mathbf{S}\mathrm{U}\neg\phi} \stackrel{\text{def}}{=} \underline{\mathbf{G}\mathrm{D}\neg\phi \;\&\; \neg\mathbf{B}\mathrm{D}\phi \;\&\; \neg\mathbf{G}\mathrm{D}\phi}$
Indifferent (**I**N)	$\mathbf{I}\mathrm{N}\phi \stackrel{\text{def}}{=} \neg\mathbf{G}\mathrm{D}\phi \;\&\; \neg\mathbf{B}\mathrm{D}\phi \;\cancel{\&\; \neg\mathbf{G}\mathrm{D}\neg\phi \;\&\; \neg\mathbf{B}\mathrm{D}\neg\phi}$
Offence of Omission (**O**F¬)	$\underline{\mathbf{O}\mathrm{F}\neg\phi} \stackrel{\text{def}}{=} \underline{\mathbf{B}\mathrm{D}\neg\phi \;\&\; \neg\mathbf{G}\mathrm{D}\phi \;\&\; \neg\mathbf{B}\mathrm{D}\phi}$
Impermissible (**I**M)	$\mathbf{I}\mathrm{M}\phi \stackrel{\text{def}}{=} \mathbf{B}\mathrm{D}\phi \;\cancel{\&\; \mathbf{G}\mathrm{D}\neg\phi}$
Offence of Commission (**O**F)	$\underline{\mathbf{O}\mathrm{F}\phi} \stackrel{\text{def}}{=} \underline{\cancel{\mathbf{B}\mathrm{D}\phi \;\&\; \neg\mathbf{G}\mathrm{D}\neg\phi} \;\&\; \neg\mathbf{B}\mathrm{D}\neg\phi}$

All indications are that in the REQ framework, the impersonally obligatory is identified with what ought to be and the personally obligatory with what ought to be brought about, but what is good is identified

with what ought to be, so the obligatory is either what is good or what is good to bring about. In either event, no reference to what is bad to omit adds anything, and a similar redundancy would result for the impermissible, as well as for the indifferent as defined above given the analysis offered for good and bad in the REQ framework. *More importantly*, the four categories of offensive commissions and omissions and of supererogatory commissions and omissions would be ruled out as incoherent *as defined above* in the 1963 framework, given the analysis offered for *good* and *bad* in the REQ framework. So this constitutes a repudiation of much (not all) of what occurred in his first framework.

In general, goodness in the REQ framework is defined as what ought to be, and badness as what ought to be absent, as T44 and T45 indicate. There is really only one special axiological notion invoked to define each here: that of *what ought to be*, a notion that intuitively invokes what is ideal in some sense. Also note that if two items are mutually exclusive, then they cannot both be good/bad in the REQ framework:

(T46) $\neg\Diamond(p \;\&\; q) \rightarrow \neg(\mathbf{G}\mathrm{D}p \;\&\; \mathbf{G}\mathrm{D}q)$
(T47) $\neg\Diamond(p \;\&\; q) \rightarrow \neg(\mathbf{B}\mathrm{D}p \;\&\; \mathbf{B}\mathrm{D}q)$

Yet I would stipulate that as a criterion of adequacy:

> Any representation of what is good must allow for the consistency of the goodness of mutually exclusive pairs; similarly, for what is bad.

For it might be good for me to help neighbor 1 all day and good for me to help neighbor 2 all day, but not possible to do both. Nor would it help to define goodness via $\neg\mathbf{O}\mathrm{U}\neg$ with the thought that if $\mathbf{O}\mathrm{U}$ picks out features of what is invariably ideal, $\neg\mathbf{O}\mathrm{U}\neg$ picks out features that are at least compatible with what is ideal. The problem then is that surely two mutually exclusive things can be both good, *and* one be better than the other, so that at most one can be compatible with what is *ideal*. I would stipulate a second criterion of adequacy:

> Any representation of what is good, must allow for the consistency of the goodness of mutually exclusive pairs, where one is more good than the other; similarly, for what is bad.

For it might be good for our mailwomen to save just Tiny Tim from the fire, but even better to save Tiny Tara too. Clearly, the representations of good and bad in this framework are too stringent.

We have seen already that some important desiderata for a framework for supererogation have been met, distinct representations of *optionality* and *indifference* (D14, D15 and D16), with the latter notion properly entailing the former (e.g. see T35), and each operator is logically indifferent to negation (T32, T36). A number of additional plausible principles governing supererogation and kindred notions follow as well. Here are two. An analog to Urmson's Criterion, strengthened to include offences (as in [Chisholm, 1963b]), follows:

(T48) $(\mathbf{SUBA}p \lor \mathbf{OFBA}p) \rightarrow (\mathbf{O}\mathrm{P}^{\mathrm{d}}p \;\&\; \neg\mathbf{INBA}p)$ (UC′ Analog)

So in particular, the indifference of bringing about p excludes its being a case of supererogation or of offence, as it should:

(T49) $\mathbf{INBA}p \rightarrow \neg(\mathbf{SUBA}p \lor \mathbf{OFBA}p)$

With D11 and D12, we have proposed analyses of supererogatory and offensive *performances*, but what of non-performances, which Chisholm always intends to include? Here there is another gap, and the fix is not at all straightforward. For example, consider this first stab:

(D19) $\mathbf{SU}\neg\mathbf{BA}\phi \stackrel{\text{def}}{=} \mathbf{O}\mathrm{U}^{\mathrm{b}}\neg\phi \;\&\; \mathbf{P}\mathrm{E}^{\mathrm{b}}\mathbf{BA}\phi \;\&\; \mathbf{P}\mathrm{E}^{\mathrm{b}}\neg\mathbf{BA}\phi$ [it is supererogatory to omit acting so that ϕ (version 1)]

A problem here is that it is supererogatory for me to not bring about ϕ does not entail in fact that it ought to be that $\neg p$:

> Cake Case: It may be that it is supererogatory for me to omit having the last piece of cake, leaving it for you, but it needn't follow from that that it ought to be that I don't have the last piece. Perhaps it is impersonally optional that I have it ($\mathbf{O}\mathrm{P}^{\mathrm{b}}i$), and likewise for your having it ($\mathbf{O}\mathrm{P}^{\mathrm{b}}u$), it only ought to be that one of us does ($\mathbf{O}\mathrm{U}^{\mathrm{b}}(i \lor u)$, but neither one of us in particular; and indeed this seems in one way to fit well with the intended analysis of supererogatory performance at least — there is something that ought to be, but it is not incumbent on me (or you) to make it so (or to omit making it so).

However, the astute reader will notice two more problematic things about D19. *The definiens is the same as that for an offensive com-*

mission, and so this follows immediately, despite its prima facie implausibility:

(T50) $\mathbf{S}\mathrm{U}\neg\mathbf{B}\mathrm{A}p \leftrightarrow \mathbf{O}\mathrm{F}\mathbf{B}\mathrm{A}p$

Secondly, T50 is essentially one of Meinong's Laws of Omission discussed above, which Chisholm explicitly was at pains to reject in [Chisholm, 1963b], and persuasively so. So D19 is an inadequate analysis of a supererogatory omission.

An alternative might be to try to retain the flavor of the analysis for a supererogatory *commission* in the framing of a supererogatory omission by treating the latter as a "supererogatory *commission of an omission*" to use Chisholm-ian language, or a supererogatory case of *refraining*, to use more contemporary language.[66] The idea would be that an omission will be supererogatory if and only if the omission itself is brought about by the agent, and it is *that* positive exercise of agency that is supererogatory:

(D20) $\mathbf{S}\mathrm{U}'\neg\mathbf{B}\mathrm{A}\phi \stackrel{\text{def}}{=} \mathbf{S}\mathrm{U}\mathbf{B}\mathrm{A}\neg\mathbf{B}\mathrm{A}\phi$ [It is supererogatory to omit acting so that ϕ (version 2)]

We thus roll the omission into the form of the original analysis of a supererogatory commission as expressed in D11. This proposal would then entail:

(T51) $\mathbf{S}\mathrm{U}'\neg\mathbf{B}\mathrm{A}p \leftrightarrow \mathbf{O}\mathrm{U}^{\mathrm{b}}\neg\mathbf{B}\mathrm{A}p \;\&\; \mathbf{P}\mathrm{E}^{\mathrm{b}}\neg\mathbf{B}\mathrm{A}\neg\mathbf{B}\mathrm{A}p \;\&\; \mathbf{P}\mathrm{E}^{\mathrm{b}}\mathbf{B}\mathrm{A}\neg\mathbf{B}\mathrm{A}p$

Paraphrasing loosely, a supererogatory omission of p is one such that it ought to be that Jane does not act so that p, even though it is both permissible for her to *not make herself* omit p and permissible for her *to make herself* omit p. What might be an example of the intended sort? Suppose because Jane was the first in her group to volunteer for (and go on) the first dangerous mission that called for a volunteer, she is now given a choice about the relative safety of the standing position she holds vis a vis the front line. Now add that because of her skills it ought to be the case that she takes a position at least intermediate to the front line, so that it ought to be that she does not pick the furthest position (*1st clause*), even though it is permissible for her to pick it and

[66] I omit many things when knocked unconscious, but do not bring it about that I omit them, and in various ways you can bring it about that I don't bring something about, where I play no part in bringing about that omission.

so permissible for her to not bring it about that she does not pick the furthest position (*2nd clause*), yet it is also permissible that she bring it about that she does not pick the furthest position, (*3rd clause*) say by picking an intermediate position or by letting the captain place her.

D20 would seem to provide no improvement over D19 in handling our cake example. For it might be supererogatory for me to omit taking the last piece — leaving it for you, even if that involves a higher order exercise of my agency (resisting the temptation to not omit it), but it need not follow that the first conjunct in the proposed definiens is true for the reasons mentioned before.

Perhaps Chisholm could just say regarding the cake example that the omission would be praiseworthy, but his analysis is not meant to capture that *agent-evaluative feature*, but something more keyed to evaluating the products of exercises of agency. As we will see in the next section, [McNamara, 2011a; McNamara, 2011b], using a different framework, argues that there is a pre-theoretic difference between going *beyond the call*, and doing so in a praiseworthy manner, and he suggests reserving the latter sort of additional agent-evaluative condition for those that are "supererogatory". Whether this will do we must set aside, noting that there appears to be a cost here. For here if the agent acts so as to not have the last piece of cake, thereby leaving it for the other person, it seems she has done more than she had to do (sacrificed more personal good than she had to and produced more for others than she had to), whatever her motives (cf. [McNamara, 2011b]).

There are obvious symmetrical analogs to the definitions above for offensive omissions:

(D21) $\mathbf{O}\mathrm{F}\neg\mathbf{B}\mathrm{A}\phi \stackrel{\text{def}}{=} \mathbf{O}\mathrm{U}^{\mathrm{b}}\phi \;\&\; \mathbf{P}\mathrm{E}^{\mathrm{b}}\neg\mathbf{B}\mathrm{A}\phi \;\&\; \mathbf{P}\mathrm{E}^{\mathrm{b}}\mathbf{B}\mathrm{A}\phi$ [it is an offence to omit acting so that ϕ (version 1)]

(D22) $\mathbf{O}\mathrm{F}' \neg\mathbf{B}\mathrm{A}\phi \stackrel{\text{def}}{=} \mathbf{O}\mathrm{F}\mathbf{B}\mathrm{A}\neg\mathbf{B}\mathrm{A}\phi$ [it is an offence to omit acting so that ϕ (version 2)]

As with D19, the definiens of D21 is one we've seen before, as this theorem indicates:

(T52) $\mathbf{O}\mathrm{F}\neg\mathbf{B}\mathrm{A}p \leftrightarrow \mathbf{S}\mathrm{U}\mathbf{B}\mathrm{A}p$,

and this is another one of the Meinongian Laws of Omission that Chisholm explicitly (and persuasively) argued against in

[Chisholm, 1963b].[67] Unpacking D22, give us this:

(T53) $\mathbf{O}\mathrm{F}'\neg\mathbf{B}\mathrm{A}p \leftrightarrow \mathbf{O}\mathrm{U}^{\mathrm{b}}\mathbf{B}\mathrm{A}p \;\&\; \mathbf{P}\mathrm{E}^{\mathrm{b}}\mathbf{B}\mathrm{A}\neg\mathbf{B}\mathrm{A}p \;\&\; \mathbf{P}\mathrm{E}^{\mathrm{b}}\neg\mathbf{B}\mathrm{A}\neg\mathbf{B}\mathrm{A}p$

What of indifference's exclusion of supererogatory commissions and omissions? These do hold for either of the two definitions in the case of omissions, whether accepting the proposed definitions that generate Laws of Omission or the second more complex analysis via commissions of omissions, as the following theorems indicate. (T54 is an analog to T49, but for the case of *commissive omissions*, and T55 and T56 each cover all cases using either $\mathbf{S}\mathrm{U}\neg$ or $\mathbf{S}\mathrm{U}'\neg$:

(T54) $\mathbf{I}\mathrm{N}\mathbf{B}\mathrm{A}p \rightarrow \neg(\mathbf{S}\mathrm{U}'\neg\mathbf{B}\mathrm{A}p \vee \mathbf{O}\mathrm{F}'\neg\mathbf{B}\mathrm{A}p)$

(T55) $\mathbf{I}\mathrm{N}\mathbf{B}\mathrm{A}p \rightarrow \neg(\mathbf{S}\mathrm{U}\mathbf{B}\mathrm{A}p \vee \mathbf{S}\mathrm{U}\neg\mathbf{B}\mathrm{A}p \vee \mathbf{O}\mathrm{F}\mathbf{B}\mathrm{A}p \vee \mathbf{O}\mathrm{F}\neg\mathbf{B}\mathrm{A}p)$

(T56) $\mathbf{I}\mathrm{N}\mathbf{B}\mathrm{A}p \rightarrow \neg(\mathbf{S}\mathrm{U}\mathbf{B}\mathrm{A}p \vee \mathbf{S}\mathrm{U}'\neg\mathbf{B}\mathrm{A}p \vee \mathbf{O}\mathrm{F}\mathbf{B}\mathrm{A}p \vee \mathbf{O}\mathrm{F}'\neg\mathbf{B}\mathrm{A}p)$

What of conflicts of supererogation? We now have two definitions of supererogatory omissions, and for supererogatory commissions, there are strong and weak version of conflicts. This framework, as we have expanded it, rules out all versions, as the fourth summary corollary indicate:

(T57) $\neg(\mathbf{S}\mathrm{U}\mathbf{B}\mathrm{A}p \;\&\; \mathbf{S}\mathrm{U}\neg\mathbf{B}\mathrm{A}p)$ [No Supererogatory Weak Conflicts]

(T58) $\neg(\mathbf{S}\mathrm{U}\mathbf{B}\mathrm{A}p \;\&\; \mathbf{S}\mathrm{U}\mathbf{B}\mathrm{A}\neg p)$ [No Supererogatory Strong Conflicts]

(T59) $\neg(\mathbf{S}\mathrm{U}\mathbf{B}\mathrm{A}p \;\&\; \mathbf{S}\mathrm{U}'\neg\mathbf{B}\mathrm{A}p)$ [No Supererogatory Weak Conflicts$'$]

(T60) $\mathbf{S}\mathrm{U}\mathbf{B}\mathrm{A}p \rightarrow \neg(\mathbf{S}\mathrm{U}\mathbf{B}\mathrm{A}\neg p \vee \mathbf{S}\mathrm{U}\neg\mathbf{B}\mathrm{A}p \vee \mathbf{S}\mathrm{U}'\neg\mathbf{B}\mathrm{A}p)$ [No Supererogatory Conflicts]

Given the symmetries regarding the corresponding concepts of offence, analogous theorems follow:

[67] As we will see later on, for quasi-supererogatory and quasi-offensive actions, analogous equivalences are more plausible.

(T61) $\neg(\mathbf{OFBA}p \ \& \ \mathbf{OF}\neg\mathbf{BA}p)$ [No Weak Conflicts of Offence]
(T62) $\neg(\mathbf{OFBA}p \ \& \ \mathbf{OFBA}\neg p)$ [No Strong Conflicts of Offence]
(T63) $\neg(\mathbf{OFBA}p \ \& \ \mathbf{OF}'\neg\mathbf{BA}p)$ [No Weak Conflicts of Offence′]
(T64) $\mathbf{OFBA}p \rightarrow \neg(\mathbf{OFBA}\neg p \vee \mathbf{OF}\neg\mathbf{BA}p \vee \mathbf{OF}'\neg\mathbf{BA}p)$ [No Conflicts of Offence]

Analogs to what we called above "Meinong's Deontic-Axiological" principles readily follow:

(T65$^{\mathrm{B1}}$) $(\mathbf{SUBA}p \vee \mathbf{OU}^{\mathrm{b}}p \vee \mathbf{OU}^{\mathrm{d}}p) \rightarrow \mathbf{G}\mathrm{D}p$ (MD-A′)
(T66$^{\mathrm{B1}}$) $(\mathbf{OFBA}p \vee \mathbf{OU}^{\mathrm{b}}\neg p \vee \mathbf{OU}^{\mathrm{d}}\neg p) \rightarrow \mathbf{B}\mathrm{D}p$ (MD-A′)

Chisholm's Threefold Axiological Classification principle follows:

(T67) $\mathrm{MJ}^3(\mathbf{G}\mathrm{D}p, \mathbf{B}\mathrm{D}p, \mathbf{N}\mathrm{U}p)$ (CTAC)

What of Meinong's Fivefold Classification? There are two possible versions, one of which fails, but the more natural one succeeds. The one that fails (with the normative positions reordered a bit) is this one:

$\mathrm{MJ}^5(\mathbf{OU}^{\mathrm{b}}p, \mathbf{OU}^{\mathrm{b}}\neg p, \mathbf{SUBA}p, \mathbf{INBA}p, \mathbf{OFBA}p)$

We label the five positions (categories) as follows (in some cases via minor equivalents cast in terms of $\mathbf{OU}^{\mathrm{b}}$):

a) $\mathbf{OU}^{\mathrm{b}}p$
b) $\mathbf{OU}^{\mathrm{b}}\neg p$
c) $\mathbf{OU}^{\mathrm{b}}p \ \& \ \neg\mathbf{OU}^{\mathrm{b}}\mathbf{BA}p \ \& \ \neg\mathbf{OU}^{\mathrm{b}}\neg\mathbf{BA}p$
d) $\neg\mathbf{OU}^{\mathrm{b}}p \ \& \ \neg\mathbf{OU}^{\mathrm{b}}\neg p \ \& \ \neg\mathbf{OU}^{\mathrm{b}}\mathbf{BA}p \ \& \ \neg\mathbf{OU}^{\mathrm{b}}\neg\mathbf{BA}p$
e) $\mathbf{OU}^{\mathrm{b}}\neg p \ \& \ \neg\mathbf{OU}^{\mathrm{b}}\neg\mathbf{BA}p \ \& \ \neg\mathbf{OU}^{\mathrm{b}}\mathbf{BA}p$

First, mutual exclusiveness fails: For example, a) conflicts with c) only if c) is itself incoherent since c) contains a)[68]. Next, consider exhaustion for a)-e). This is equivalent to if neither a) nor b), then at least one of c)-e); but a) and b) both being false rules out both c) and e) immediately, so must d) be true? No, in particular, $\neg\mathbf{OU}^{\mathrm{b}}\neg\mathbf{BA}p$ does not follow, nor should it. It may be impersonally optional whether p or $\neg p$, but it might

[68]So c) is a real category only if there is no conflict.

be impersonally obligatory that I not bring p about, as it is outside my jurisdiction.

The version that is more symmetrical and natural results from shifting a) and b) above ($\mathbf{O}\mathrm{U}^{\mathrm{b}}p$; $\mathbf{O}\mathrm{U}^{\mathrm{b}}\neg p$) to focus on Jane Doe's agency or not regarding p, as these are proposed as representing *personal* obligation and prohibition,

a′) $\mathbf{O}\mathrm{U}^{\mathrm{b}}\mathbf{B}\mathrm{A}p$
b′) $\mathbf{O}\mathrm{U}^{\mathrm{b}}\neg\mathbf{B}\mathrm{A}p$

Leaving c)-e) unchanged, the result in this version:

(T68) $\mathrm{MJ}^5(\mathbf{O}\mathrm{U}^{\mathrm{b}}\mathbf{B}\mathrm{A}p, \mathbf{O}\mathrm{U}^{\mathrm{b}}\neg\mathbf{B}\mathrm{A}p, \mathbf{S}\mathrm{U}\mathbf{B}\mathrm{A}p, \mathbf{I}\mathrm{N}\mathbf{B}\mathrm{A}p, \mathbf{O}\mathrm{F}\mathbf{B}\mathrm{A}p)$

This is indeed derivable, which is a plus in as much as a central point in this area is to find an expanded partition of the (obligatory, prohibited and) optional. However, the exhaustion component implies that anything neither obligatory, impermissible nor indifferent to bring about is either supererogatory or offensive to bring about. This is dubious, so not really a plus to derive. As we will see in Section 5, sometimes doing even the minimum required can be good and admirable and optional since one can do even more, but it will not thereby be an offence (and can't be beyond the call if it's the minimum).

Finally, what of a REQ-based analog of C7FC (Chisholm-like Seven-Fold Classification)?
There are four possible versions to assess, but in the interest of space and time, we zero in on the most plausible two renderings, where agency is robustly present.

We have two versions of the seven disjunctions depending on how we interpret supererogatory and offensive omissions again. The first version is this one:

(T69) $\mathrm{MJ}^7(\mathbf{O}\mathrm{U}^{\mathrm{b}}\mathbf{B}\mathrm{A}p, \mathbf{O}\mathrm{U}^{\mathrm{b}}\neg\mathbf{B}\mathrm{A}p, \mathbf{S}\mathrm{U}\mathbf{B}\mathrm{A}p, \mathbf{S}\mathrm{U}\neg\mathbf{B}\mathrm{A}p, \mathbf{I}\mathrm{N}\mathbf{B}\mathrm{A}p, \mathbf{O}\mathrm{F}\mathbf{B}\mathrm{A}p, \mathbf{O}\mathrm{F}\neg\mathbf{B}\mathrm{A}p)$

Which is indeed provable, but it is a *trivial* corollary of T68. For as we noted in T50 and T52, the fourth and seventh positions are logically equivalent to the sixth and third positions respectively, as is reflected again by comparing the defining conditions of the conjuncts defining conditions of the fourth with the sixth, and of the seventh with the third. So there is no real expansion on this interpretation beyond the second of the five-fold partitions above. Thus shifting to a′) and b′) on

this analysis of supererogatory and offensive omissions just takes us to the second version of the *five*-fold classification. So let's consider the version that makes agency more prominent in the fourth and seventh disjuncts:

$\mathrm{MJ}^{7}(\mathbf{O}\mathrm{U}^{\mathrm{b}}\mathbf{B}\mathrm{A}p,\ \mathbf{O}\mathrm{U}^{\mathrm{b}}\neg\mathbf{B}\mathrm{A}p,\ \mathbf{S}\mathrm{U}\mathbf{B}\mathrm{A}p,\ \mathbf{S}\mathrm{U}'\neg\mathbf{B}\mathrm{A}p,\ \mathbf{I}\mathrm{N}\mathbf{B}\mathrm{A}p,$
$\mathbf{O}\mathrm{F}\mathbf{B}\mathrm{A}p,\ \mathbf{O}\mathrm{F}'\neg\mathbf{B}\mathrm{A}p)$

For convenience, we list the seven defining conditions cast via $\mathbf{O}\mathrm{U}^{\mathrm{b}}$:

a′) $\mathbf{O}\mathrm{U}^{\mathrm{b}}\mathbf{B}\mathrm{A}p$
b′) $\mathbf{O}\mathrm{U}^{\mathrm{b}}\neg\mathbf{B}\mathrm{A}p$
c) $\mathbf{O}\mathrm{U}^{\mathrm{b}}p\ \&\ \neg\mathbf{O}\mathrm{U}^{\mathrm{b}}\mathbf{B}\mathrm{A}p\ \&\ \neg\mathbf{O}\mathrm{U}^{\mathrm{b}}\neg\mathbf{B}\mathrm{A}p$
d) $\neg\mathbf{O}\mathrm{U}^{\mathrm{b}}p\ \&\ \neg\mathbf{O}\mathrm{U}^{\mathrm{b}}\neg p\ \&\ \neg\mathbf{O}\mathrm{U}^{\mathrm{b}}\mathbf{B}\mathrm{A}p\ \&\ \neg\mathbf{O}\mathrm{U}^{\mathrm{b}}\neg\mathbf{B}\mathrm{A}p$
e) $\mathbf{O}\mathrm{U}^{\mathrm{b}}\neg p\ \&\ \neg\mathbf{O}\mathrm{U}^{\mathrm{b}}\neg\mathbf{B}\mathrm{A}p\ \&\ \neg\mathbf{O}\mathrm{U}^{\mathrm{b}}\mathbf{B}\mathrm{A}p.$
f) $\mathbf{O}\mathrm{U}^{\mathrm{b}}\neg\mathbf{B}\mathrm{A}p\ \&\ \neg\mathbf{O}\mathrm{U}^{\mathrm{b}}\mathbf{B}\mathrm{A}\neg\mathbf{B}\mathrm{A}p\ \&\ \neg\mathbf{O}\mathrm{U}^{\mathrm{b}}\neg\mathbf{B}\mathrm{A}\neg\mathbf{B}\mathrm{A}p$
g) $\mathbf{O}\mathrm{U}^{\mathrm{b}}\mathbf{B}\mathrm{A}p\ \&\ \neg\mathbf{O}\mathrm{U}^{\mathrm{b}}\neg\mathbf{B}\mathrm{A}\neg\mathbf{B}\mathrm{A}p\ \&\ \neg\mathbf{O}\mathrm{U}^{\mathrm{b}}\mathbf{B}\mathrm{A}\neg\mathbf{B}\mathrm{A}p$

Here we run into a difficulty that raises questions about the adequacy of the analysis of the associated concepts: the attempt to prove mutual exclusivity generates a dilemma. Working backwards, f) rules out g) by T11; e) rules out f) and g) tautologically; d) rules out e)-g) tautologically, c) rules out d)-g) either tautologically or by T11 in the case of e). This leave a′) and b′). a′) conflicts with c)-e) tautologically and with b) and the first conjunct of f) via T11, leaving g), whose first conjunct is a′), and whose third conjunct is implied by a′), for $\mathbf{O}\mathrm{U}^{\mathrm{b}}\mathbf{B}\mathrm{A}\neg\mathbf{B}\mathrm{A}p \rightarrow \mathbf{O}\mathrm{U}^{\mathrm{b}}\neg\mathbf{B}\mathrm{A}p$ by $\mathrm{T23}^{\mathrm{B1}}$ and D8 but a′) rules out $\mathbf{O}\mathrm{U}^{\mathrm{b}}\neg\mathbf{B}\mathrm{A}p$ given T11; so the only remaining source of possible conflict is the second conjunct, and for these to conflict, it would have to be a theorem that

($\mathrm{T70}^{\mathrm{B3}}$) $\mathbf{O}\mathrm{U}^{\mathrm{b}}\mathbf{B}\mathrm{A}p \rightarrow \mathbf{O}\mathrm{U}^{\mathrm{b}}\neg\mathbf{B}\mathrm{A}\neg\mathbf{B}\mathrm{A}p,$

which formula is derivable given B3, $p\mathrm{R}\mathbf{B}\mathrm{A}q \rightarrow p\mathrm{R}\neg\mathbf{B}\mathrm{A}\neg\mathbf{B}\mathrm{A}q$, which seems plausible, but it does again involve invoking another **B**A-R linking axiom not considered by Chisholm (none are). Let's assume T70 is sound *for sake of argument*. Now comes the dilemma: T70 is sound only if g′) is *incoherent*, for its first conjunct is a′), the antecedent of, T70, but its second conjunct is the denial of the consequent of T70. So we get this result:

($\mathrm{T71}^{\mathrm{B3}}$) $\neg\mathbf{O}\mathrm{F}'\neg\mathbf{B}\mathrm{A}p$, *for any p*

The problem, cast another way, is that we get what appeared to be the most plausible representation of the analog to the sevenfold classification from the prior two frameworks (RCGB and [Chisholm and Sosa, 1966a]) only to find that we can't prove mutual exclusion of the first disjunct with the sixth (supererogatory omission) without rendering the seventh (offensive omission) incoherent and empty, so in either event, we can't get a significant version of the sevenfold classification.

Let's take stock.

4.5 Evaluation of the REQ framework for supererogation and kin

The REQ framework is impressive, especially in terms of its expressive resources. I have developed the implications of the framework quite a bit beyond what was present, but in doing so, I have stuck close to the spirit and intent of the original. The shear variety of target concepts for which analyses are proposed is impressive, not to mention that Chisholm first proposed the framework in 1964. But here we must reflect on some challenges the framework faces, some general or foundational in the framework, some more specific to the focus of this chapter. Let's begin with some aspects of the general framework.

Chisholm does not say a lot about the fundamental relation of requirement, but he does give some alternative glosses, "calls for", "apt", "fitting", as potential stand-ins for "requires", as well as indicating that the relations scope can extend to relations between colors, musical chords/keys, and figures. But I think the equivalence of "requires" with "fitting" that Chisholm suggest, "p could be said to be fitting to q provided q requires p" is doubtful. *Requires* is a strong term, and fits better with strong deontic terms like makes *obligatory*, makes it a *duty*, makes it *prohibited*, makes it a *must*. Saying something is *apt* or *fitting* seems to say something much weaker and fits better with terms like *good, valuable, sensible* or even *ideal*. This is more than a quibble about words. $\mathbf{O}\mathrm{U}^{\mathrm{b}}p$ is meant to analyze both what ought to be, as well as what is impersonally obligatory, and likewise $\mathbf{O}\mathrm{U}^{\mathrm{b}}\mathbf{B}\mathrm{A}p$ is meant to analyze both what the assumed agent ought to bring about and what it is obligatory for the agent to bring about. But it can't do both these things. If I say that "it must not be that the children are left to starve", I say something much stronger (and more appropriate given the content –children starving) than "it ought not be that children are left to starve". Likewise, if we say "Jane must feed her children", we say something much stronger and more apt than merely saying "Jane ought to feed her children". It

is the stronger notion that is needed for what is obligatory, permissible, prohibited, optional, and such. But now there is a problem. We want a theory that can represent both what must be done and what ought to be done, and what one must do, and what one ought to do, and, of course, distinctly within both pairs.[69] So we cannot read $\mathbf{O}\mathrm{U}^{\mathrm{b}}p$ as "it ought to be that p" but must instead read it as something like "it is mandatory that p" or "it must be that p".[70] We have already pointed out the problem with the analysis of good (and of bad) offered when we do read $\mathbf{O}\mathrm{U}^{\mathrm{b}}$ as "it ought to be that", but when we read it as "it is mandatory that" then matters with the analysis of good and bad become worse surely (e.g. now something is good iff mandatory). Indifference is also inadequately characterized. We saw formal indications of problems, but now we have an additional more substantive problem. Some p might be such that it is not mandatory that it be the case nor that it not be the case nor that Jane brings it about nor that Jane does not do so, and yet still surely be something that *ought to be* the case, and/or that Jane *ought to do*, or alternatively it can be something that *ought not be*, and that Jane *ought not do*. So something can be supererogatory or offensive and yet be indifferent on the proposed analysis, once we adjust the reading of $\mathbf{O}\mathrm{U}^{\mathrm{b}}p$, as I think we must, so that it can have a chance of representing what is obligatory, impermissible, optional, etc. The resources are just too limited. A single primitive, be it interpreted via the stronger, "requires" or via the weaker, "fitting", will not be enough to properly analyze the target concepts. We also saw in discussing what ought to be and what is good, that a proper account of these two must allow for two incompatible things to be good, and one more good than the other and so at least one such that it ought *not* be if the other ought to be. I think this sort of structure of differentially ranked permissible options is essential to action beyond the call.

Furthermore, our cake example was a problem for all versions of supererogatory omissions we explored, and it did not look like there was anything in the framework to allow something being supererogatory to bring about (or to avoid), even though it is false that it ought to be the case. The analysis can at best only cover a subset of the targeted cases.

[69] It is beyond the scope of this chapter to explore the introduction of two such relations, and a development that would parallel some aspects of Chisholm's framework

[70] Chisholm himself gives as a gloss for supererogation "you ought to but you don't have to" [Chisholm, 1964, p. 152], but, although the distinction is insightful and ahead of its time, he does not see the full significance of the distinction in the context of the REQ framework.

We saw that there was a substantial unexplored gap at the level of linking agency to the requirement primitive, and thus to the derived notions, and potential axioms that have a plausible ring. We introduced a simple logic for an agency operator, $\mathbf{B}$A, and then four tentative axioms, B1-B4, linking requirement to agency. In places, the links are needed to generate *plausible* results (e.g. $\mathbf{O}\mathrm{U}^{\mathrm{b}}\mathbf{B}\mathrm{A}p \rightarrow \mathbf{O}\mathrm{U}^{\mathrm{b}}\neg\mathbf{B}\mathrm{A}\neg p$), and also in order to generate some of the desired linkages between supererogation and kindred notions that there is reason to think Chisholm wished to endorse. But it is also the case that some of these links render some of his analyses *redundant*, and we saw in trying to generate the sevenfold classification expected, we generate one desired mutual exclusion of two categories only at the expense of rendering another category incoherent. Furthermore, as the astute reader may have noticed, B2, generates an undesirable consequence:

(T72$^{\mathrm{B2}}$) $\neg(\mathbf{S}\mathrm{U}\neg\mathbf{B}\mathrm{A}p \vee \mathbf{O}\mathrm{F}\mathbf{B}\mathrm{A}p)$, for any p

For by T10$^{\mathrm{B2}}$, $\mathbf{P}\mathrm{E}^{\mathrm{b}}\mathbf{B}\mathrm{A}p \rightarrow\mathbf{P}\mathrm{E}^{\mathrm{b}}p$, that is, $\mathbf{P}\mathrm{E}^{\mathrm{b}}\mathbf{B}\mathrm{A}p \rightarrow \neg\mathbf{O}\mathrm{U}^{\mathrm{b}}\neg p$ via D9, but the first two conjuncts of the definiens of $\mathbf{S}\mathrm{U}\neg$ are $\mathbf{O}\mathrm{U}^{\mathrm{b}}\neg p$ & $\mathbf{P}\mathrm{E}^{\mathrm{b}}\mathbf{B}\mathrm{A}p$, and this is not just another problem for an independently problematical analysis of supererogatory omissions, for we also saw that this way of construing a supererogatory omission is exactly how Chisholm analyzed an offence, so adding B2 renders that notion, involving positive agency (not omissions), incoherent. Thus aside from the fact that this way of characterizing a supererogatory omission sanctions Meinong's implausible laws of omission, B2 independently rules out offensive commissions. Thus although Chisholm clearly intended his account to apply to commissions and omissions, he overlooked the challenges often associated with accounting for omissions per se, and for their normative status, as well as overlooking the need to more carefully articulate how requirement is to be linked to agency, and we proposed four natural links:

(B1) $p\mathrm{R}\mathbf{B}\mathrm{A}q \rightarrow p\mathrm{R}q$
(B2) $p\mathrm{R}\neg q \rightarrow p\mathrm{R}\neg\mathbf{B}\mathrm{A}q$
(B3) $p\mathrm{R}\mathbf{B}\mathrm{A}q \rightarrow p\mathrm{R}\neg\mathbf{B}\mathrm{A}\neg\mathbf{B}\mathrm{A}q$
(B4) $p\mathrm{R}\mathbf{B}\mathrm{A}\neg q \rightarrow p\mathrm{R}\neg\mathbf{B}\mathrm{A}q$

Unlike B2 which links a required proposition's absence with a required absence of the production of said proposition by any agent, the other three formulae link states of affairs requiring oneÕs positive agency with

other requirements. However, B2 is plausible in its own right, and without it we would then lose $\mathbf{P}\mathrm{E}^{\mathrm{b}}\mathbf{B}\mathrm{A}p \rightarrow \mathbf{P}\mathrm{E}^{\mathrm{b}}p$ ($\mathrm{T20}^{\mathrm{B2}}$), which on its intended reading is quite plausible. But we have already indicated as well that some of these potential axioms are disruptive, and less importantly, if added, they would generate some redundancies in the definitions This leaves one wondering if this would be welcome or if it indicates we are overlooking some way of conceiving the relation of requirement that makes some of B1-B4 implausible. But then what relationships would be plausible? No relationship at all between what ought to be and what ought to be done seems to be an implausible stance and unattractive regarding the theory's potential applications.

Furthermore, there are basic challenges faced by the Meinong-Chisholm reduction of an agent's obligation to what ought to be the case,[71] but beyond those it is essential to the REQ approach that something can be impersonally mandatory without being obligatory for any particular agent, which is a plus, as it might be mandatory that someone in the department do something, but not mandatory that any one particular individual do so. This is a strength of the framework, and fits well with at least many cases of volunteering, where we have a situation where not only is it the case that $\mathbf{O}\mathrm{U}^{\mathrm{b}}p$ and $\mathbf{O}\mathrm{P}\mathbf{B}\mathrm{A}_j p \vee \mathbf{O}\mathrm{P}\mathbf{B}\mathrm{A}_s p$), where we have two agents ($j \neq s$), but also $\mathbf{O}\mathrm{U}^{\mathrm{b}}(\mathbf{B}\mathrm{A}_j p \vee \mathbf{B}\mathrm{A}_s p)$. Communities and groups often fit this bill, and thrive only because someone steps up and does what is needed. But when we turned to supererogatory and offensive *omissions*, we ran into special difficulties in getting a representation of these that seemed to fit in with the spirit of the account while retaining the essential classificatory linkages. We also saw with the cake example that there can be a supererogatory omission on my part of the last piece of cake (thus leaving it to equally-situated you), but where it is not the case that it ought to be that I omit the cake, thus not seeming to fit the proposed analysis. Furthermore, in the cake case, it seems to come out as supererogatory to bring it about that one of the agents eats the last slice by simply doing so, which is implausible.

There is much left unexplored here in what is already lengthy coverage of Chisholm's investigations in the 1960s and 1970s into areas relevant to our chapter topic (not to mention from the same period, his seminal [Chisholm, 1963a]. His work (including that with Sosa) constitutes, far and away, the most substantial sustained attempt before the 1990s by a single author to sort out the conceptual neighborhood that *supererogation* belongs to. His work is full of insights, and has been

[71] For example, see the discussion of objections in [Horty, 2001].

highly influential, despite, ironically, not having been given the scrutiny it deserves.

5 Doing Well Enough (DWE)

Introduction

In a variety of places, McNamara has articulated and defended the first model-theoretic framework designed to represent supererogation and kindred notions, called "DWE", for "Doing Well Enough". Here we provide an introduction to the core framework he developed in [McNamara, 1990; McNamara, 1996b; McNamara, 1996a; McNamara, 1996c], as well as in [Mares and McNamara, 1997]. We do this by providing a series of independent pathways to the core model-theoretic framework, which McNamara takes to provide a cumulative case that the framework is generally on the right track, as well as to indicate that the diverse pathways are each intimately important to understanding the conceptual neighborhood of supererogation.[72] In the process, he explores often-neglected concepts or distinctions, as well as mistaken presuppositions that have pervaded both ethical theory and deontic logic, and have substantially increased the difficulty of finding a place for supererogation, as well as resulting in a comparative paucity of expressive resources in ethical theorizing as well as in deontic logic. The concepts in focus in the core framework are for some key moral statuses that exercises of one's agency might have. Such exercises might be *permissible, impermissible, obligatory, omissible, optional, indifferent, significant, beyond the call of duty, the least you can do, suboptimal.* He contends that these notions of common sense morality are of substantial interest in ethical theory, and that this is one (not the only to be sure) clear place where deontic logic, despite its historically marginalized status among ethical theorists, can be of some service.

[72]The following framing of the works specified above has been included in presentations at various recent venues: "Toward a More Fine-Grained Conceptual Scheme for Moral Statuses". Keynote Address. 12th International Conference on Deontic Logic and Normative Systems (DEON 2014), Ghent University, Ghent, Belgium, July 2014; "Toward a Taxonomical Framework for Some Fundamental Moral Concepts", University of Southern California, Deontic Modality Workshop, May 22, 2013; "Toward a Formal Framework for Some Fundamental Common Moral Statuses", Munich Center for Mathematical Philosophy, Formal Ethics I, Ludwig Maximilian University, Munich, October 11, 2012. I thank the members of the audiences for their helpful comments, especially, Fenrong Liu, Marek Sergot, Janice Dowell, Ralph Wedgwood, Angelika Kratzer, and John Broome.

5.1 Indifference and optionality

Conflation of optionality with indifference

As already indicated in Section 2 of this chapter to say that an exercise of agency is optional (**O**P) is to say that it is neither obligatory (**O**B) nor impermissible (**I**M), but to say of such an exercise that it is a matter of indifference is to say something much stronger. Yet there has been a pervasive conflation in 20th century ethical theory and in deontic logic of these two concepts, typically by standardly reading "neither obligatory nor impermissible" as "indifferent" (**I**N). We find this in G. E. Moore, Von Wright, Prior, even in Urmson as he struggles to distinguish the two, and one continues to regularly find the condition of being neither obligatory nor impermissible read as *indifference*. This leads immediately to what we called "Moral Rigor" (MR) in Section 2.1: $\mathbf{I}\mathrm{N}\phi \leftrightarrow \mathbf{O}\mathrm{P}\phi$, which rules out going beyond the call of duty, since $\mathbf{B}\mathrm{C}\phi \rightarrow (\mathbf{O}\mathrm{P}\phi \;\&\; \neg\mathbf{I}\mathrm{N}\phi)$. The left to right direction of MR is fine, but not the right to left, and this is essentially Urmson's Constraint (UC) on deontic schemes: $\mathbf{I}\mathrm{N}\phi \rightarrow \mathbf{O}\mathrm{P}\phi$, but not $\mathbf{O}\mathrm{P}\phi \rightarrow \mathbf{I}\mathrm{N}\phi$.

Semantic frameworks for optionality and indifference

Consider first a simple classical framework for optionality and non-optionality (**N**O):

$\vDash_i \mathbf{O}\mathrm{P}\phi$: ϕ holds in some i-acceptable world (A^i) and $\neg\phi$ holds in some i-acceptable world.

$\vDash_i \mathbf{N}\mathrm{O}\phi$: either ϕ holds in all i-acceptable world or $\neg\phi$ holds in all i-acceptable world

Where A^i represents the set of worlds acceptable from i, we can represent these conditions as in Figure 15.

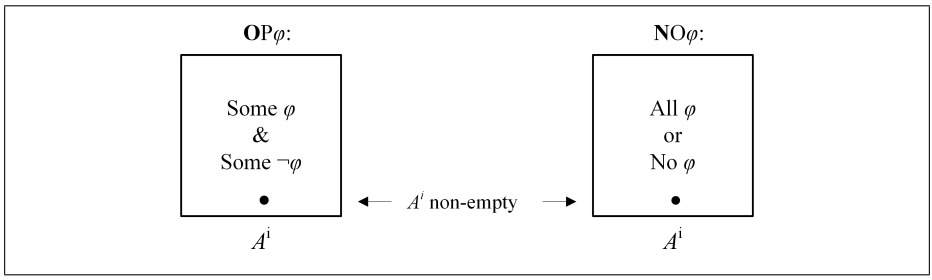


Figure 15: semantics for **O**P and **N**O

McNamara then proposes this alternative non-standard semantic structure and analysis for indifference:

$\models_i$ **I**Nϕ: for every i-level of value, there is a ϕ-world at that level, and a $\neg\phi$-world at that level.

$\models_i$ **SI**ϕ: in some i-level of value, all worlds there are ϕ-worlds or all the worlds there are $\neg\phi$-worlds.

We can picture things as in Figure 16.

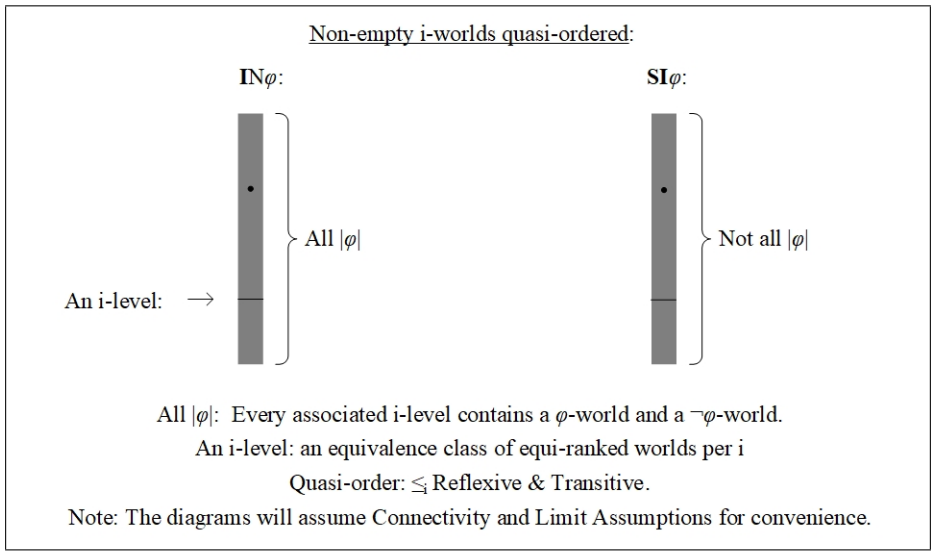


Figure 16: Semantics for **I**N and **SI**

That is, we assume a reflexive and transitive ordering relation on a non-empty set of worlds, and then roughly, ϕ is a matter of indifference just in case for every level of worlds, there is a ϕ-world and also a $\neg\phi$-world at that level–every level of value can be achieved with or without ϕ. McNamara refers to this as "deliberative indifference".[73] Conversely, something is a matter of moral significance when some level of value can only be achieved with ϕ or can only be achieved without ϕ.[74]

[73]So **I**N is not intended to capture the notion that no matter of moral value is involved, but instead to capture the idea that holistically, if no matter what level of value you can achieve in a choice situation, ϕ's presence as well as ϕ's absence is consistent with that level, then deliberation about $\phi/\neg\phi$ is idle.

[74]As noted in the diagram, neither the limit assumption nor the assumption of connectivity suggested in the diagrams is essential. See for example Section 5.6 below, which summarizes results from [Mares and McNamara, 1997], where determination theorems are given for systems without either assumption.

Relating the two operators: compression vs. augmentation

Semantically, there are two natural ways to think about the relationship between **O**P logics and **I**N logics:

a. *Compression*: Stipulate no more than one i-level for each world, and the **I**N/**SI** logic is indistinguishable from that for **O**P/**N**O (See Figure 17).

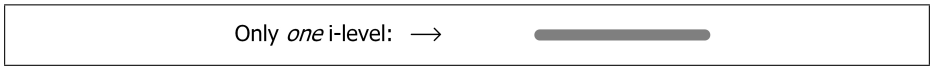


Figure 17: One-level compression

Here the contingency/optionality logic is seen as a special case of the indifference logic, the logic determined by collapsing the i-levels to just one, but there is a more expansive way of thinking of the two *together*.

b. *Augmentation*: Suppose we wish to represent **O**P and **I**N in one unified system? We might then stipulate this: For each i in the **I**N frames, select a non-empty upper subset of i-levels: A^i = the *i-acceptable* levels/cells. See Figure 18. An *i-acceptable world* is any an upper region world. Then represent $\models_i$ **O**Pϕ and $\models_i$ **I**Nϕ as before.

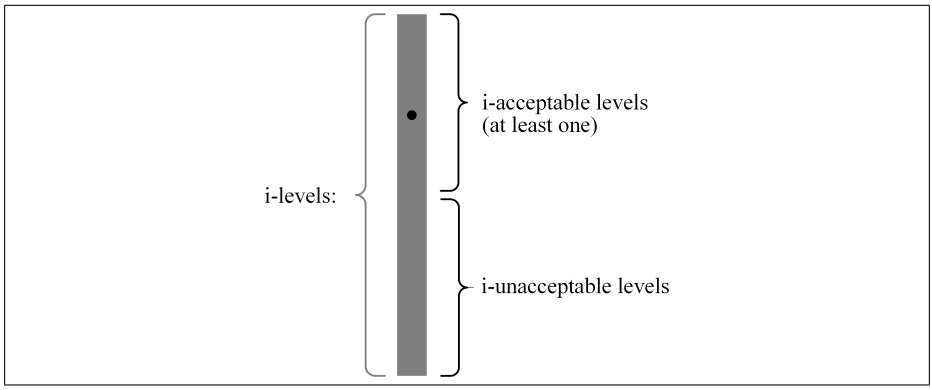


Figure 18: Augmented semantic structure

Clearly $\models$ **I**N$\phi \rightarrow$ **O**Pϕ (given $A^i \neq \varnothing$), for if every level has a ϕ-world and a $\neg\phi$-world, and there are acceptable worlds, then

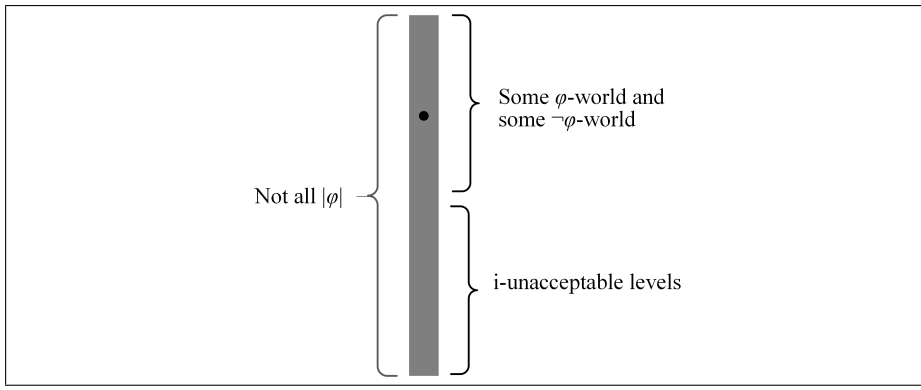


Figure 19: Optionality without indifference

there is at least one where ϕ holds and another where it does not. However, $\not\models \mathbf{OP}\phi \rightarrow \mathbf{IN}\phi$, as the model in Figure 19 indicates.

So, happily, Urmson's Constraint is satisfied by the framework we just arrived at by thinking about indifference and optionality in tandem.

5.2 *Must* and *ought*

The pervasive conflation of must and ought, a bipartisan presupposition

In "Must I Do What I Ought? (or Will the Least I can Do Do?)", McNamara argues that

> A community of scholars might mistake an expression to be continuous with some concept of philosophical concern that it is in fact not continuous with, and they might do so largely unreflectively, with all the risks of potential confusion that unexamined assumptions can typically engender. I would like to suggest that just such an assumption has pervaded ethical theory and deontic logic this century. And it is this assumption that I will argue is mistaken–by arguing for a negative answer to the main title question. [McNamara, 1996c, p. 154]

Against the conflation on its face. [McNamara, 1996c], and in more detail, [McNamara, 1994], and Chapter 3 of [McNamara, 1990], provide a cumulative case argument that *must* properly entails *ought*,

and that it is *must*, not *ought* that has the traditional relationships to permissibility and impermissibility and obligatoriness. He offers the first model-theoretic account of the difference between *must* and *ought*. Here we briefly summarize.

Firstly, McNamara asks us to consider two groups of expressions, where the two groupings seem natural, and the members of the first group seem uniformly stronger than those of the second group:

S *must* see to it that ϕ.
S has to see to it that ϕ.
S is obligated to see to it that ϕ.
S is required to see to it that ϕ.
It is S's duty to see to it that ϕ.
It is imperative that S see to it that ϕ.
It is incumbent on S to see to it that ϕ.

S *ought* to see to it that ϕ.
S should see to it that ϕ.
It is morally advisable that S see to it that ϕ.
It is morally preferable that S see to it that ϕ.
It is morally best that S sees to it that ϕ.
It is morally most appropriate that S see to it that ϕ.
It is morally ideal that S see to it that ϕ.

Yet members of the first group have been routinely conflated with those of the second group, often used interchangeably. This interchangeability can be seen as something like a bipartisan presupposition of much of 20th century ethical theory and deontic logic.[75] Focusing on *must* and *ought* from our two groups, there is a *prima facie intuitive difference in strength*. It looks like *must* properly entails *ought*: for example, what one must do, one ought to do, but not necessarily vice versa.

Conversational differences: ought conversationally implies optional Secondly, there are *conversational differences*. If I say you ought to take this exit, it suggests that this exit is best, but it also suggests that there are other acceptable though less good options, and if there were none, you might rightly complain and say "Why didn't you say that I must take that exit!?" This conversational implication is difficult to explain if *must* and *ought* are semantically equivalent; but it

[75] A few examples: [Moore, 1912 (1965 edition), p. 15]; [von Wright, 1951, p. 58]; [von Wright, 1963, p.73 and p.83].

is easy to explain, in a very familiar way, if *must* properly implies *ought*. By using "ought", your listener, assuming you are in the know, infers that the equally accessible "must" does not apply, so one ought to turn at this exit, but it is not the case that you must do so, and thus that it is ok to not turn, and so it is optional, even if turning is preferable. If you must turn here, there are no acceptable alternatives. So *ought's* conversational implication of optionality supports *must's* semantic entailment of *ought*. Note also that although the example is morally neutral, the same phenomenon is reflected in moral contexts and exchanges, as [McNamara, 1994], and Chapter 3 of [McNamara, 1990] note. For example. A: "Well, I suppose I ought to go to the meeting". B: "What do you mean 'you *ought* to go'?! You *must* go, period! You're the one that demanded there be a meeting in the first place."

Constitutional differences: deadlines Thirdly, there are *constitutional differences*. A *deadline* is a time by which something must be done, not one by which something ought to be done. *Ought* is too weak. Similarly, a job *requirement* is something that must be done, not one that merely ought to be done, as illustrated by the widespread: "Employees must wash their hands."

Speech act differences: commanding vs. recommending Fourthly, there are *illocutionary / speech act differences* between what we do with *must* and *ought*. If your employer tells you that you *must* do some difficult task, she is typically *commanding* you to do it, but if she tells you that you *ought* to do it, she is typically *recommending* or *advising* you to do it. Ignoring the latter might cost you a raise, ignoring the former might cost you your job.

Contrastive claims: you ought to, but you don't have to Fifthly, there are *contrastive differences*. "You ought to but you don't have to"[76] (or "You ought to but it is not the case that you must") seems perfectly apt, whereas "You ought to but it is not the case that you ought to" clearly is not apt, and "You must do so, but you ought not do so" also seems incoherent without some special story.[77] Similar

[76] Chisholm insightfully notices the relevance of this to supererogation in [Chisholm, 1963b], although he seems to take it to be coextensive with going beyond the call, which is doubtful. We will return to the latter later.

[77] For example, one that puts the "must do so" in scare quotes, for example according to some rule or law deemed unjust or unfit, or one that has the "must do so" refer to some compulsion on the addressee's part that is to be resisted.

remarks pertain to expressions like "You can, but you ought not."

Can, can't and must vs. can, can't and ought Sixthly, there is *pressure from interactions with "can" and "can't".* To say "you can turn here" is to say "it is not the case that you must not turn here", to say "you can't turn here" is to say "you mustn't turn here" and to say "you must turn here" is to say you "can't not turn here". But if we substitute "ought" for must in these they lose their prima facie plausibility. As McNamara [1990, 1994 1996c] argues in more detail, whereas

$$\begin{array}{l} \text{a) } \mathrm{CAN}\phi \leftrightarrow \neg\mathrm{MUST}\neg\phi \\ \text{b) } \mathrm{MUST}\neg\phi \leftrightarrow \mathrm{CANT}\phi \\ \text{c) } \mathrm{MUST}\phi \leftrightarrow \neg\mathrm{CAN}\neg\phi \end{array}$$

hold, when we substitute "ought" for "must" above, although the three right to left implications hold, the left to right implications fail. Add to this that "can" and "can't" clearly can and do routinely express *permissibility* and *impermissibility* and are thus continuous with traditional concerns with these notions, and an obvious moral follows:

> "Must", but not "ought", expresses whatever ethicists and deontic logicians have virtually uniformly taken "ought" to express: *moral or deontic necessity.* For the latter has routinely been taken to be whatever satisfies the familiar definitional equivalences involving permissibility and impermissibility. And this means that, contrary to a dominant bipartisan trend this [now past] century, we can't take "ought" as basic and then assume that what is permissible is whatever satisfies "∼ought∼", nor that what is impermissible is whatever satisfies "ought∼"... [McNamara, 1996c, p.158]

Perhaps after only "good", "ought" has been the most studied expression in 20th century ethical theory.[78] A nice representative statement is this one:

> The Two Main ethical concepts are expressed respectively by the words "good" and "ought" (or "duty") ..... The action that we ought to do is also called our "duty" [Ewing, 1953, p. 12 and p. 15].

[78] Going back to [Moore, 1903].

This presupposed that *ought* had the tight continuity with the traditional concerns with what is obligatory, permissible and impermissible often expressed in a deontic square of opposition, but it does not.[79]

Pressure from the use of modals in other domains Seventhly, there is *pressure from other domains* where these modal auxiliaries are used. It makes perfect sense to say that based on the evidence about the deck and the past cards appearing "The next card ought to be a spade, but it need not be, though it must be a spade or a club." In epistemic contexts it is plain that *must* is stronger than *ought*. This puts additional pressure on acknowledging their difference in deontic and ethical contexts. Indeed, McNamara argues for this in more detail elsewhere [McNamara, 1990; McNamara, 1994] (and in passing in [Mares and McNamara, 1997]), citing what he calls a "Field Invariance" hypothesis about the relationship between "must", "ought", "can", and "can't" and close cousins: that there implicational relationships are generally invariant across domains where their use is felicitous.[80]

Contexts where strong modals are felicitous, but "ought" is not Eighthly, consider that there are contexts where "must"/"have", but not "ought" are felicitous, like law. Why? Well "must", "can", and "can't" are felicitous and widely used because these indicate what is mandatory, permitted, and prohibited according to the laws. However, "ought" is generally out of place. Why? Aside from the fact that using it would conversationally imply *optionality*, which would be a disaster for stating a legal *requirement*, I think it indicates something else: the absence of a relevant ordering. Law tells us what is acceptable, unacceptable, and mandatory, but it does not provide any suitable ranking of options that could get it into the business of making widespread pronouncements about what is best, but not mandatory. This tends to confirm that, at least in practical contexts, *ought* is tied to a ranking of alternatives. [McNamara, 1994], and Chapter 3 of [McNamara, 1990] make a similar point regarding alethic modal contexts and close cousins of "must", "can", "can't", and the absence of "ought" or any close cousin to it.

[79] It is in this sense that a tacit pervasive but mistaken presupposition of 20th Century ethical theory (and of deontic logic) has been that *must* is equivalent to *ought* in ethical contexts.

[80] Note that this makes a weaker claim, and so in a sense, a safer claim, than that the terms are univocal.

The bipartisan presupposition and the marginalization of supererogation Now notice the impact of the pervasive conflation in 20th century ethical theory and deontic logic of deontic necessity with *ought*. Consider the near axiomatic claim: "You ought to do the best you can". By implication of the conflation, we get "You must do the best you can", but then how can the best you can do ever be supererogatory? Similarly, if what you ought to do, you must do, and so you can't not do, then nothing supererogatory can ever be something you ought to do, not even small favors. So once again, we have a conflation that makes it difficult to find a stable or coherent place for supererogation.

A conservative framework for *must* and *ought*

Here is a conservative response to these reflections. Reject the equivalence of must and ought in deontic contexts, but retain the tie of ought to what is best (in some sense of "best"). A natural simple semantic structure results then from mapping *must* to what is done at all acceptable worlds, and ought as what is done at the best of those, retaining a common ordering of the worlds:

$\vDash_i$ **M**Uϕ: every i-acceptable world is a ϕ-world
$\vDash_i$ **O**Uϕ: all the i-best (or i-best i-acceptable worlds) are ϕ-worlds

Figure 20 provides a diagrammatic expression of the truth conditions.

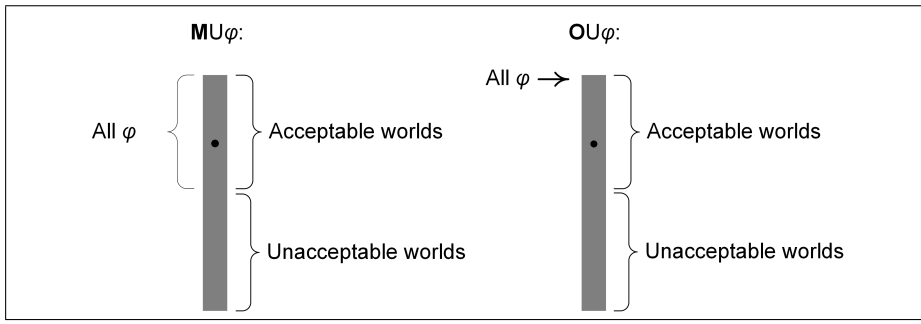


Figure 20: Semantics for **M**U and **O**U

This semantic picture can be used to explain all the data cited earlier to motivate the non-equivalence of *must* and *ought* (see the references above), and obviously *must* properly entails *ought* in this picture. We arrive independently at a structure much like the one for indifference and optionality (Figure 21).

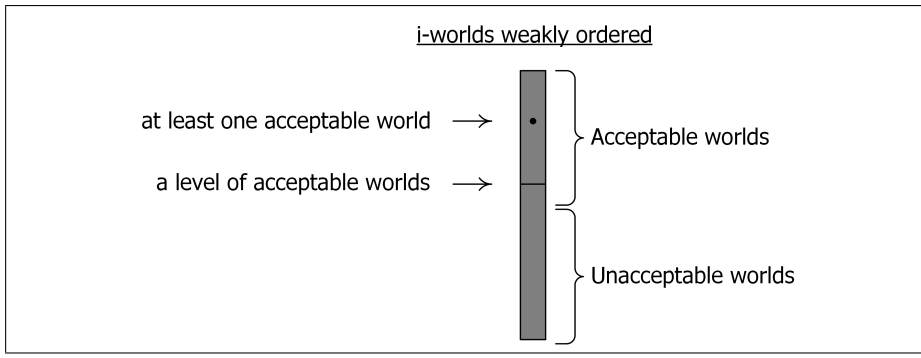


Figure 21: The same semantic structures emerge

Levels emerge naturally again in the frames, so that a framework for full indifference emerges naturally too, and unmotivated by any reflections on "Indifference". Conversely, as we saw earlier, augmenting our **I**N frames to represent **O**P naturally led to the same structures.

5.3 An ignored construction: *the least you can do*

This idiom has received virtually no attention in deontic logic or ethical theory, yet it is ripe with import. Consider "*The least you can do is call and let them know you won't show*". The "can" appears to be the *can* of permissibility.[81] The idiom suggests an *ordering* quite naturally: an ordering with a *minimally acceptable level*, and lower and higher levels potentially, and with all lower levels being impermissible. The latter is reflected in the aptness of this idiom when used to scold: "The least you could have done was called" with its contextually understood, "and you didn't even do that much!" This is an extremely rich idiom and an important data point in understanding pre-theoretic moral consciousness. We are once again naturally led to the same structures, where the least one can do is mapped to the minimal acceptable level of one's alternatives:

$\models_i \mathbf{L}\mathrm{E}\phi$: all the lowest ranked i-acceptable worlds are ϕ-worlds

We can picture this two ways, focusing on the i-acceptable worlds only, or on all the worlds divided into acceptable and unacceptable ones (Figure 22).

[81] It can't be the can of ability or possibility plainly, and there seem no plausible candidates given the use of this but for the *can* of permissibility.

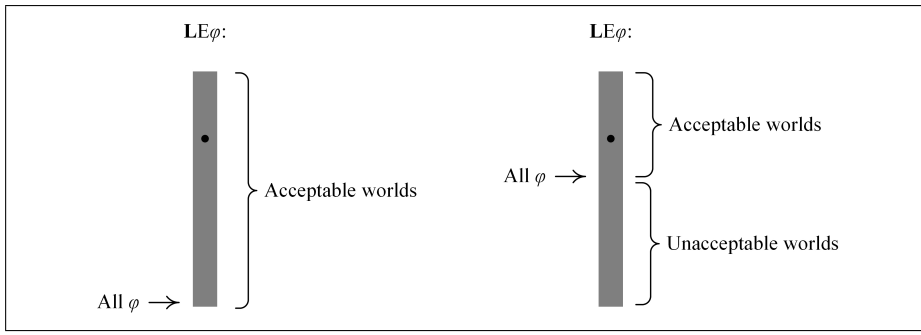


Figure 22: Semantics for **L**E

We have essentially the same structures again, generated now by reflecting on "the least one can do".

5.4 *Doing more (good) than you have to do*

This form of speech is quite colloquial. The "more" (like "least") suggests an *ordering*, and since it is more than you had to do, it suggests the acceptability of doing less, so *ordered acceptable options* naturally emerge. Also, it naturally suggests the possibility of doing less than you had to do. So we are on the way to the same sort of structures again. Consider this condition: *someone does more than she would have done had she done the least she could have done.* As a first stab, McNamara suggests $\mathbf{B}\mathrm{C}\phi \stackrel{\text{def}}{=} \mathbf{P}\mathrm{E}\phi \;\&\; \mathbf{L}\mathrm{E}\neg\phi$ (it is beyond the call that ϕ iff ϕ *is permissible but precluded by the least one can do*). Semantically, this means $\neg\phi$ holds at the lowest ranked acceptable worlds, but ϕ holds at some acceptable world.

$\vDash_i \mathbf{B}\mathrm{C}\phi$: all the lowest ranked i-acceptable worlds are $\neg\phi$-worlds, but some acceptable world is a ϕ-world

Figure 23 provides a diagrammatic expression of the truth conditions.

Note that $\vDash \mathbf{B}\mathrm{C}\phi \rightarrow (\neg\mathbf{I}\mathrm{N}\phi \;\&\; \mathbf{O}\mathrm{P}\phi)$. This is generated by the definition and truth conditions, in keeping with the main motivation for Urmson's proposed Constraint ($\nvDash \mathbf{O}\mathrm{P}\phi \rightarrow \mathbf{I}\mathrm{N}\phi$). Once again, a familiar structure emerges.

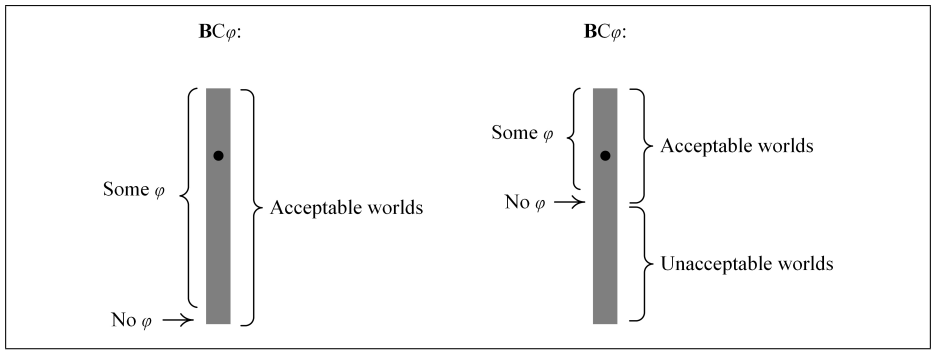


Figure 23: Semantics for **B**C

5.5 *"You ought to but don't have to"; "you can but ought not"*

Reflecting on these expressions, and assuming we find linking "ought" to "best" plausible, we are naturally lead to positing the same structures by reasoning similar to that for doing more than you must. Also, given the above interpretations of *must/have to* and *ought*, we also naturally get a mirror image operator (it is *permissibly suboptimal* that): **PS**ϕ: $\stackrel{\text{def}}{=}$ **O**U$\neg\phi$ & **P**Eϕ.

$\models_i$ **PS**ϕ: all the highest ranked i-acceptable worlds are ϕ-worlds, but some acceptable world is a $\neg\phi$-world.

Figure 24 provides a diagrammatic expression of the truth-conditions.

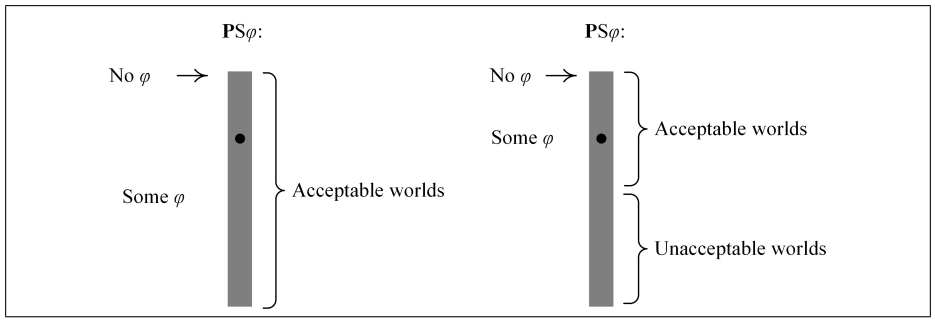


Figure 24: Semantics for **P**S

5.6 Upshot: a cumulative case for logical and semantic framework

Summing up the DWE framework

Take any individual reflection point from Sections 5.1-5.5, and you can motivate the same structures. We have interlocking support for DWE greater than the sum of the evidential value of the parts. We can review the framework by considering the following case. Imagine that you have to provide some delicate information to a colleague across campus, and for simplicity, imagine there are three ways to do this, by emailing, phoning or talking in person. Lastly, suppose, not implausibly, that the permissible options are ranked according to how personal they are. So giving the info by email is the lowest ranked of the permissible options, giving the info by phone is next best, and giving the info in person is the best way (Figure 25).

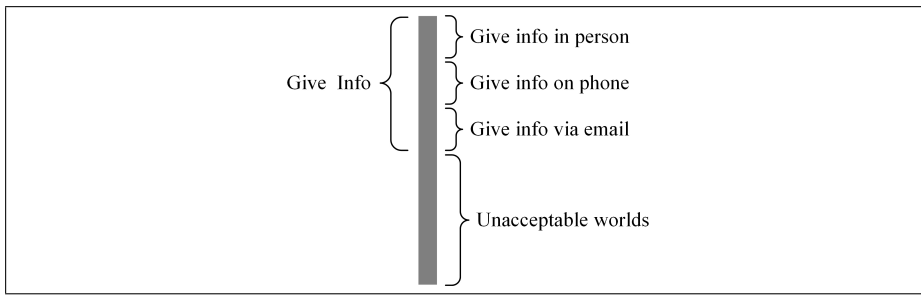


Figure 25: Semantic framework review

Then, not giving the info is impermissible, and giving the info is something you must do. The least you can do is provide the info by email, giving it in person is best, and so it is what you ought to do. However, giving the info by calling on the phone is not only permissible but beyond the call, as is doing it in person, furthermore, giving the info via email or by phone is permissibly suboptimal.

Below are displays of the conditions for the familiar five operators, and the new ones McNamara proposed (Figures 26 and 27).

The increase in expressive power and complexity is not marginal. Recall the Traditional Threefold Classification. DWE generates the following analog to the TTC in Figure 28, one where we move to a twelvefold partition. All the main action is within the optional sphere, as the external annotation indicates. Recall also the Traditional Deontic Square. A merely partial analog in DWE to the traditional deontic square (or

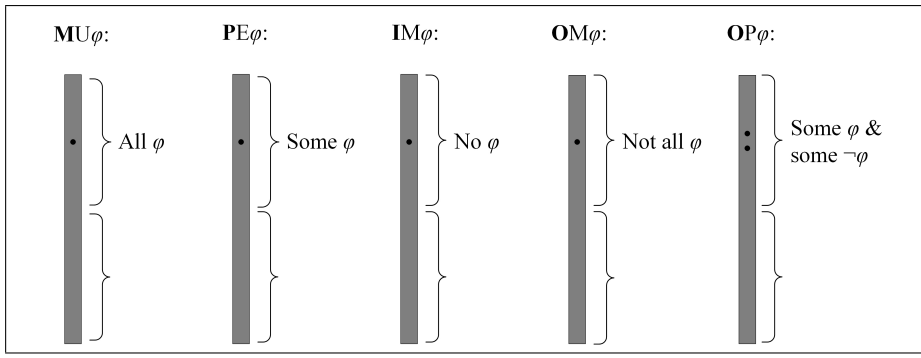


Figure 26: Semantics for the standard operators

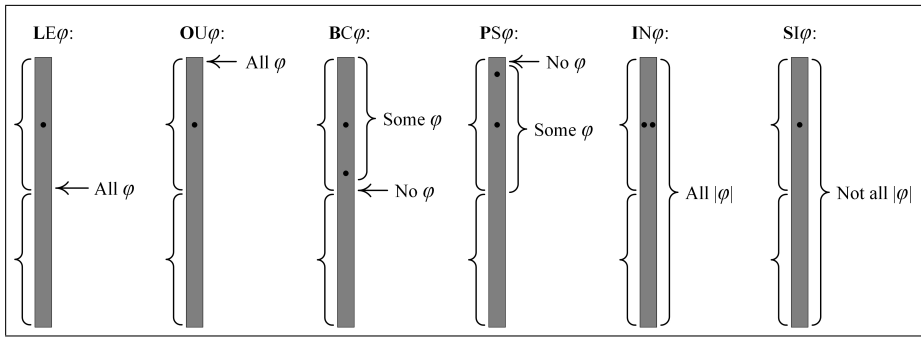


Figure 27: Semantics for the new operators

hexagon) is the octodecagon in Figure 29.[82]

Syntax, semantics, determination

With one minor deviation (to be discussed later) in the representation of indifference, the framework sketched above can be regimented as follows. The four primitive operators for the core DWE framework are these:

[82]Quite partial. For example, consider just the operator **O**U. Neither $\neg$**O**Uϕ nor $\neg$**O**U$\neg\phi$ nor $\neg$**O**Uϕ & $\neg$**O**U$\neg\phi$ (optimality indifference) are listed on any of the nodes, and likewise for the **L**E, **P**S, **B**C operators. Thus, there is only *one subcontrary* relation indicated. A truer analog would require a thirty-sided regular polygon (triacontagon).

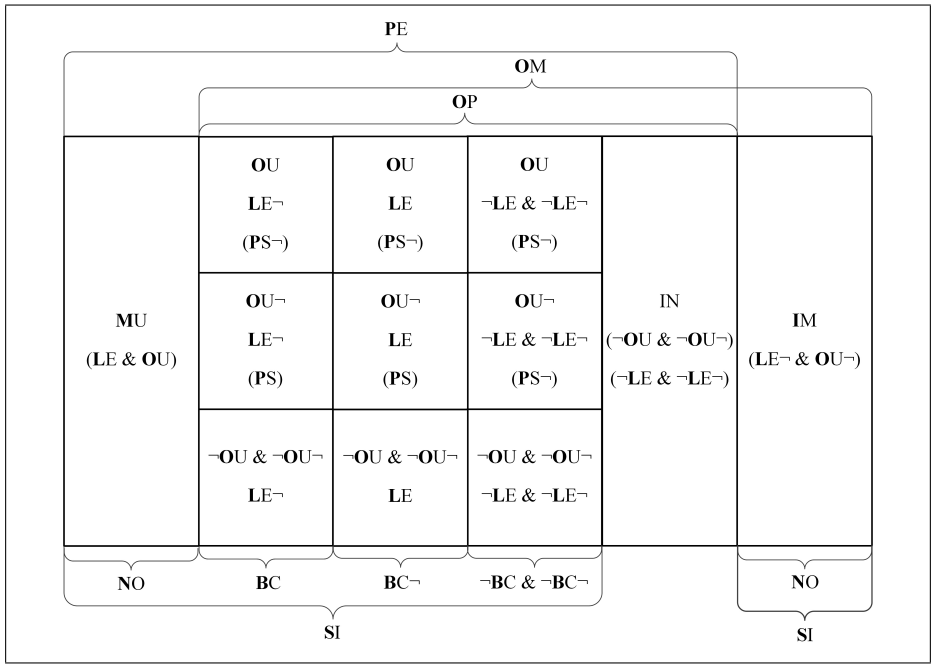


Figure 28: The DWE twelve-fold classification

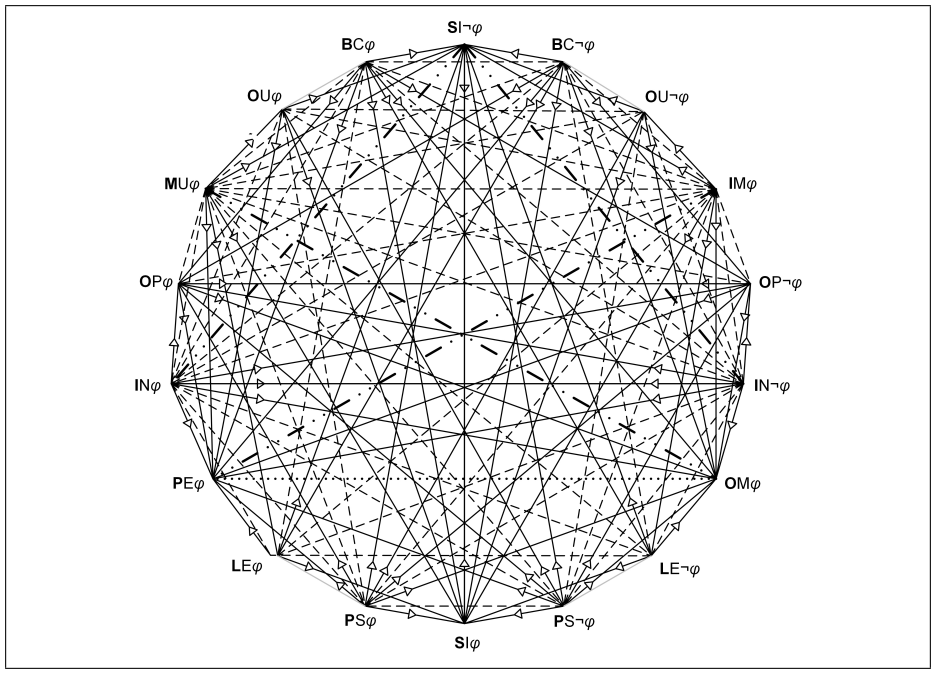


Figure 29: A DWE octodecagon

$\mathbf{O}\mathrm{B}\phi$:	It is *Obligatory* (for S) that ϕ
$\mathbf{MI}\phi$:	The *Minimum* (for S) involves/implies (its being the case that) ϕ
$\mathbf{MA}\phi$:	The *Maximum* (for S) involves/implies (its being the case that) ϕ.
$\mathbf{I}\mathrm{N}\phi$:	It is (fully) *Indifferent* (for S) that ϕ[83]

We imagine these operators added to some language for classical propositional logic, and taking any formula as argument. Some defined operators, and their intended readings can then be introduced:

$\mathbf{P}\mathrm{E}\phi \stackrel{\text{def}}{=} \neg\mathbf{O}\mathrm{B}\neg\phi$.	(It is *Permissible* for S that ϕ.)
$\mathbf{I}\mathrm{M}\phi \stackrel{\text{def}}{=} \mathbf{O}\mathrm{B}\neg\phi$.	(It is *Impermissible* for S that ϕ.)
$\mathbf{O}\mathrm{M}\phi \stackrel{\text{def}}{=} \neg\mathbf{O}\mathrm{B}\phi$.	(It is Omissible (for S) that ϕ
$\mathbf{O}\mathrm{P}\phi \stackrel{\text{def}}{=} \neg\mathbf{O}\mathrm{B}\phi \;\&\; \neg\mathbf{O}\mathrm{B}\neg\phi$.	(It is *Optional* for S that ϕ.)
$\mathbf{S}\mathrm{I}\phi \stackrel{\text{def}}{=} \neg\mathbf{I}\mathrm{N}\phi$.	(It is *Significant* for S that ϕ.)
$\mathbf{B}\mathrm{C}\phi \stackrel{\text{def}}{=} \mathbf{P}\mathrm{E}\phi \;\&\; \mathbf{MI}\neg\phi$.	(It is *Beyond the Call* for S that ϕ.)
$\mathbf{P}\mathrm{S}\phi \stackrel{\text{def}}{=} \mathbf{P}\mathrm{E}\phi \;\&\; \mathbf{MA}\neg\phi$.	(It is *Permissibly Suboptimal* for S that ϕ.)
$\mathbf{OI}\phi \stackrel{\text{def}}{=} \neg\mathbf{MA}\phi \;\&\; \neg\mathbf{MI}\neg\phi$.	(It is Optimality Indifferent for S that ϕ.)
$\mathbf{MI}\phi \stackrel{\text{def}}{=} \neg\mathbf{MI}\phi \;\&\; \neg\mathbf{MI}\neg\phi$.	(It is Minimality Indifferent for S that ϕ.)
$\mathbf{PI}\phi \stackrel{\text{def}}{=} \mathbf{OI}\phi \;\&\; \mathbf{MI}\phi$.	(It is Polarity Indifferent for S that ϕ.)

[Mares and McNamara, 1997] presents two logics, DWE and $\mathrm{DWE}^{\mathrm{m}}$ a weakening of DWE. The DWE Logic is the following one, where "*" ranges over **O**B, **MA**, **MI**:

(A0)	All tautologous DWE formulas;
(A1)	$*(\phi \to \psi) \to (*\phi \to *\psi)$
(A2)	$\mathbf{O}\mathrm{B}\phi \to (\mathbf{MI}\phi \;\&\; \mathbf{MA}\phi)$
(A3)	$(\mathbf{MI}\phi \vee \mathbf{MA}\phi) \to \mathbf{P}\mathrm{E}\phi$
(A4)	$\mathbf{I}\mathrm{N}\phi \to \mathbf{I}\mathrm{N}\neg\phi$
(A5)	$\mathbf{I}\mathrm{N}\phi \to (\neg\mathbf{MI}\phi \;\&\; \neg\mathbf{MA}\phi)$
(A6)	$(\mathbf{O}\mathrm{B}(\phi \to \psi) \;\&\; \mathbf{O}\mathrm{B}(\psi \to \chi) \;\&\; \mathbf{I}\mathrm{N}\phi \;\&\; \mathbf{I}\mathrm{N}\chi) \to \mathbf{I}\mathrm{N}\psi$
(R1)	If $\vdash \phi$ and $\vdash \phi \to \psi$ then $\vdash \psi$
(R2)	If $\vdash \phi$, then $\vdash \mathbf{O}\mathrm{B}\phi$.

The semantics for the DWE Logic is as follows:

[83] Note that the readings are *personal* but not *agential* — a bit more on this below.

Let F = (W, A, ≤), be a DWE frame, where:

1. $W \neq \varnothing$
2. $A \subseteq W^2$ and $\forall i \exists$jAij (seriality)
3. $\leq \subseteq W^3$:
   (a) $(k \leq_i j$ or $j \leq_i k)$ iff (Aij & Aik), for any $i, j, k \in W$
   (b) if $j \leq_i k$ and $k \leq_i l$ then $j \leq_i l$, for any i, j, k, l in W.

P is an assignment on F: $F = (W, A, \leq$ is a DWE frame and P is a function from the propositional variables (PV) to Power(W), defined on PV.

$M = (F, P)$ is a DWE model: $F = (W, A, \leq)$ is a DWE frame and P is an assignment on F.

Truth at a world in a DWE model: Let $M = (F, P)$ be a DWE model, where $F = (W, A, \leq)$ and $j =_i k \stackrel{\text{def}}{=} j \leq_i k$ & $k \leq_i j$. Then for any $i \in W$, we have these truth clauses:

Basic truth-conditions at a world i, in a model, M:

[PC]	(Usual conditions for sentence letters and connectives)	
[**O**B]	$M \vDash_i$ **O**Bϕ:	$\forall j$(if Aij then $M \vDash_j \phi$).
[**M**A]	$M \vDash_i$ **M**Aϕ:	$\exists j(Aij$ & $(\forall k)$(if $j \leq_i k$ then $M \vDash_k \phi$)).
[**M**I]	$M \vDash_i$ **M**Iϕ:	$\exists j(Aij$ & $(\forall k)$(if $k \leq_i j$ then $M \vDash_k \phi$)).
[**I**N]	$M \vDash_i$ **I**Nϕ:	$\forall j$[if Aij then $\exists k(k =_i j$ & $M \vDash_k \phi)$ & $\exists k(k =_i j$ & $M \vDash_k \neg\phi)$].

Note the truth conditions for **M**A and **M**I are such that **M**Aϕ and/or **M**Iϕ can be true even if there is no limit on the ordering at either the upper end or lower end of the i-acceptable worlds. In our informal exposition above, our diagrams suggested a limit at each end. However, the evaluation of **I**N is confined to the i-acceptable range. This is a restriction compared to our informal representation, in that it is a somewhat weaker condition for indifference than one that says ϕ and $\neg\phi$ must appear among *all* i-levels of worlds including i-unacceptable ones.

Derivative truth conditions:

[**P**E] $M \vDash_i$ **P**Eϕ: $\exists j(Aij \ \& \ M \vDash_j \phi)$.

[**I**M] $M \vDash_i$ **I**Mϕ: $\forall j$(if Aij then $M \vDash_j \neg\phi$).

[**O**M] $M \vDash_i$ **O**Mϕ: $\exists j(Aij \ \& \ M \vDash_j \neg\phi)$.

[**O**P] $M \vDash_i$ **O**Pϕ: $\exists j(Aij \ \& \ M \vDash_j \phi)$ and $\exists j(Aij \ \& \ M \vDash_j \neg\phi)$.

[**S**I] $M \vDash_i$ **S**Iϕ: $\exists j[Aij$ & either $\forall k$(if $k =_i j$ then $M \vDash_k \phi$) or $\forall k$(if $k =_i j$ then $M \vDash_k \neg\phi)]$.

[**B**C] $M \vDash_i$ **B**Cϕ: $\exists j(Aij \ \& \ M \vDash_j \phi) \ \& \ \exists j[Aij \ \& \ \forall k$(if $k \leq_i j$ then $M \vDash_k \neg\phi)]$.

[**P**S] $M \vDash_i$ **P**Sϕ: $\exists j(Aij \ \& \ M \vDash_j \phi) \ \& \ \exists j[Aij \ \& \ \forall k$(if $j \leq_i k$ then $M \vDash_k \neg\phi)]$.

[**O**I] $M \vDash_i$ **O**Iϕ: $\neg\exists j(Aij \ \& \ \forall k$(if $j \leq_i k$ then $M \vDash_k \phi)) \ \& \ \neg\exists j(Aij \ \& \ \forall k$(if $j \leq_i k$, then $M \vDash_k \neg\phi))$.

[**M**I] $M \vDash_i$ **M**Iϕ: $\neg\exists j(Aij \ \& \ \forall k$(if $k \leq_i j$ then $M \vDash_k \phi)) \ \& \ \neg\exists j(Aij \ \& \ \forall k$(if $k \leq_i j$ then $M \vDash_k \neg\phi))$.

[**P**I] $M \vDash_I$ **P**Iϕ: $\neg\exists j(Aij \ \& \ \forall k$(if $j \leq_i k$ then $M \vDash_k \phi)) \ \& \ \neg\exists j(Aij \ \& \ \forall k$(if $j \leq_i k$ then $M \vDash_k \neg\phi)) \ \& \ \neg\exists j(Aij \ \& \ \forall k$(if $k \leq_i j$ then $M \vDash_k \phi)) \ \& \ \neg\exists j(Aij \ \& \ \forall k$(if $k \leq_i j$ then $M \vDash_k \neg\phi))$.

Truth in a DWE model: $M \vDash \phi$ iff $M \vDash_i \phi$, for every i in W of M.

Validity for a set of DWE models C: $\vDash \phi$ iff $M \vDash \phi$, for all M *in* C.

Metatheorem: The DWE logic is determined by the class of all DWE models.[84]

For DWE$^{\mathrm{m}}$, the only change in the logic is that we replace A1 of DWE, $^*(\phi \to \psi) \to (*\phi \to *\psi)$, with something that A1 properly entails:

(A1′) $\quad$ **M**U$(\phi \to \psi) \to (^*\phi \to {}^*\psi)$

Although DWE$^{\mathrm{m}}$ still generates a full SDL fragment for **M**U, it does not for **M**A nor for **M**I, since no conflict principles for **M**A and **M**I are no longer derivable, nor are aggregation principles:

[84] [Mares and McNamara, 1997].

$\vdash (\mathbf{MU}\phi \;\&\; \mathbf{MU}\psi) \rightarrow \mathbf{MU}(\phi \;\&\; \psi)$
$\nvdash (\mathbf{MA}\phi \;\&\; \mathbf{MA}\psi) \rightarrow \mathbf{MA}(\phi \;\&\; \psi)$
$\nvdash (\mathbf{MI}\phi \;\&\; \mathbf{MI}\psi) \rightarrow \mathbf{MI}(\phi \;\&\; \psi)$

$\vdash \mathbf{MU}\phi \rightarrow \neg\mathbf{MU}\neg\psi$
$\nvdash \mathbf{MA}\phi \rightarrow \neg\mathbf{MA}\neg\psi$
$\nvdash \mathbf{MI}\phi \rightarrow \neg\mathbf{MI}\neg\psi$

In fact, if we add the aggregations principles to $\mathrm{DWE}^{\mathrm{m}}$, the system is equipollent with DWE.

At the semantic level, we need to retract connectivity for $\leq_i$ in the frames, thereby allowing incomparable i-acceptable worlds in the frames. We merely weaken clause (3) a) in the definition of DWE frames as follows:

$3'$. $\leq \subseteq W^3$: a′) $(k \leq_i j$ only if $Aij \;\&\; Aik)$ and if Aij then $j \leq_i j$, for any $i, j \in W$

The rest is as before.

Metatheorem: The $\mathrm{DWE}^{\mathrm{m}}$ logic is determined by the class of $\mathrm{DWE}^{\mathrm{m}}$ models[85]

5.7 DWE operators: personal non-agential readings

There is no representation of agency or action in the core DWE framework. This is a limit, and furthermore, it raises questions about whether or not the operators can represent their intended target concepts, and how to read the operators (or what they can be taken to represent).[86] We address the operator reading question first, beginning with **O**B, and drawing on [McNamara, 2004]. We have often been reading "**O**B" as personal not agential, following a suggestion at the end of [Krogh and Herrestad, 1996]. "OBϕ" is intended to express a *personal* obligation, one that Jane Doe, our mock person, has, but one that does not require that Jane Doe be *the agent of*ϕ, nor that ϕ itself be a proposition asserting Doe's agency regarding some ψ. McNamara offers a provisional argument for non-agential personal obligations:

[85] [Mares and McNamara, 1997].

[86] For example, are they implicitly read agentially so that they are composites of sorts, where it is unclear what $\mathbf{O}\mathrm{B}\neg\phi$ is saying — is it denying agency regarding some ϕ or is it denying ϕ, and similarly for other operators?

> If all my obligations are agential, then each of my obligations is an obligation for *me* to bring about some thing. If each of my obligations is an obligation for me to bring about some thing, then none of my obligations can be fulfilled by someone else. But some of my obligations can be fulfilled by someone else. Therefore, not all my obligations are agential. [McNamara, 2004]

It may be obligatory for you that your child does her homework (or is fed), but she may do it on her own with no intervention at all from you. The obligation is fulfilled, though not by you, so how can that be so if the obligation was for you to bring it about? Compare a friend paying your debt. Furthermore, often I'm obligated to be in my office. This typically requires me to do things to make it so, but the order of explanation is from the obligation to be a certain way to a derived obligation to do certain things to achieve it. We are often obligated to be such that ϕ, where ϕ is not agential, and it can even happen that someone else make it so that we are such that ϕ. Being obligated to do something is not what makes an obligation personal on this account. What makes such an obligation personal is that I am responsible if the obligatory state is not realized — the buck stops at my desk as it were; in contrast, what makes an obligation strictly agential is that only the agent can fulfill it — what is obligatory is that *you, yourself, do some thing.*[87] Some, but not all, of our obligations, are like this. Now consider the following:

1) I'm obligated to be in my office.

This is an obligation on me to *be in* a location, not *to do* something. The sentential complement is non-agential. 1) can be aptly paraphrased as:

1′) $\overbrace{\text{I'm obligated to be}}^{\text{Personal}}$ such $\overbrace{\text{that I am in my office.}}^{\text{Non-Agential}}$

Then an obligation to bring it about that ϕ, an agential obligation, might be conceived as just a special case of a personal obligation to be:

[87] Perhaps put another way, I am obligated to pay my bill is not strictly agential. Rather we specify the obligation by a default, but if you pay the bill on my behalf unbeknownst to me, my obligation is met in full. Strictly, what is obligatory for me is that my bill is paid.

2) I'm obligated to be such that I bring it about that ϕ.

Note that it also seems apt to say that a person is obligated to be cooperative, just, faithful, honest, punctual, that is, to be such that she possesses the traits in question. If I have such obligations, I may fulfill some at least effortlessly in the sense that I am naturally such that I possess the trait. McNamara also points out that other evaluatives clearly take such ways to be as complements, for example, Jane can be blameworthy/praiseworthy for being stubborn. If these reflections are correct, then it may be that agency and obligation are not as tightly linked as sometimes thought.[88] It is merely a contingent fact that the vast majority of one's obligations require some exercise of one's own agency to be fulfilled.

Can we extend this strategy to other operators in the DWE framework? I think we can. Consider the following:

MU/**O**Bϕ:	Jane Doe must be such that ϕ/ It is obligatory for Doe to be such that ϕ
PEϕ:	It is permissible for Doe to be such that ϕ/ Jane Doe can be such that ϕ
IMϕ:	It is impermissible for Doe to be such that ϕ/ Jane Doe can't be such that ϕ
OMϕ:	It is omissible (non-obligatory) for Doe to be such that ϕ
OPϕ:	It is optional for Doe to be such that ϕ
OU/**M**Aϕ:	Jane Doe ought to be such that ϕ
LE/**M**Iϕ:	The least Doe can be is such that ϕ
BCϕ:	It is beyond the call for Jane Doe to be such that ϕ
PSϕ:	It is permissibly suboptimal for Jane Doe to be such ϕ
INϕ:	It is indifferent for Jane Doe to be such that ϕ
SIϕ:	It is significant for Jane Doe to be such that ϕ

The non-agential phrasings my not be the most typical, but they are coherent readings. For example, it may be that in a given situation, the least Jane Doe can be is contrite[89], and in another situation, it might be beyond the call for her to be forgiving or to be merciful, just as in another it might be a matter of indifference for her to be sleepy. But

[88] For example, see "The Restricted Complement Thesis" in [Belnap *et al.*, 2001].

[89] Notice that even "the least you can do is be contrite", despite the presence of "do", can't really be read as itself introducing any action. What is called for here is to *now* be contrite (at a minimum).

being contrite, forgiving, merciful, and sleepy are not agential, since they are states, not exercises of agency, even if there are typical agential manifestations of such states. We conclude that there is a reasonable case in favor of a non-agential yet personal reading of our operators, and that given this reading, all that is required to introduce agency in the complement is to have ϕ be agential, so that, for example, "the least Jane can do is apologize" can be recast as "the least Jane Doe can be is such that she apologizes" (or such that she brings it about that she does). Stilted, but coherent.

Let me also note here that neglecting these personal *ought/must/least* state-like constructions is risky. For example, some have argued for a restricted complement thesis to the effect that *ought*, if personal, is well-formed only if it takes an agential complement. But this seems clearly wrong. Also, contrastivists that stress actions in analyzing personal *oughts* should not overlook constructions like *Jane ought to be contrite*, which seems personal but does not derive necessarily from any *action* that is best. A set of states would seem to be required, not actions.

Although there is great interest in agency and supererogation, we postpone further exploration until later. We take this section to have shown that there is that there is at least a plausible case to be made for regimented reading of the operators as personal but non-agential, and for weaving in a separate agency operator that can be introduced into the complement of the personal but non-agential operator to generate an agential normative position.

5.8 Interlude: revisiting the DWE frame structure

Here we briefly return to reflecting on the acceptable and unacceptable worlds & their ordering since there is a puzzle and the attentive reader will have noticed a discrepancy between McNamara's formal presentation of the structures (technically, frames) in the DWE logics, and the informal multipoint motivation for the DWE framework. Recall some things:

1) The semantics and determination theorems for DWE confine the $\leq_i$ ordering relations to *i-acceptable* worlds.
2) However, the intuitive picture of the structures in our cumulative case argument had the *i*-unacceptable worlds too, and one ordering to rule them all.
3) If we think of this as due to a singular ranking, what determines the cut off?
4) Note also that the interpretation of $\mathbf{O}\mathrm{U}\phi$ amounts to ϕ holds at the best of the *i-acceptable* worlds, not simply ϕ holds at the best (*i-accessible*) worlds.
5) Here is just one related question: Are all worlds ranked as high as an *i*-acceptable world, themselves *i*-acceptable?

Let me take some steps in the direction of a clarification. It seems we must have two factors, one determining *i-acceptability* from the *i-accessible*(or *reachable*, so to speak) worlds, one determining some *i*-relative ordering of the *i-acceptable* worlds (Figure 30).

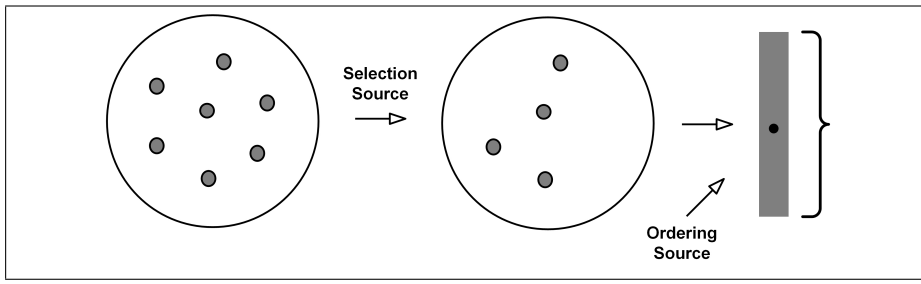


Figure 30: From accessible worlds to ordered acceptable worlds

I've also suggested elsewhere that we can think of the first as itself resulting from a partition of the *i*-accessible worlds into *at least* two trivial "levels", the *i*-acceptables on one level, the *i*-unacceptables at another.[90] A question: do we rank the *i*-unacceptable worlds in the same way as the acceptable ones — using the same ranking consideration/s? Let's work our way to an answer, setting aside the missing i-unacceptable worlds for the moment. Let's suppose that we have a non-trivial ordering source yielding the *i*-acceptable worlds from the *i-accessible/reachable/available* ones (Figure 31).
For sake of concreteness: assume the ranking above of the accessible/reachable worlds is by justice. How might we now rank the *i*-acceptable worlds? Not by a justice ranking. They are all tied at the

[90] Originally in [McNamara, 1988], but see also [McNamara, 1990] and [McNamara, 1996b].

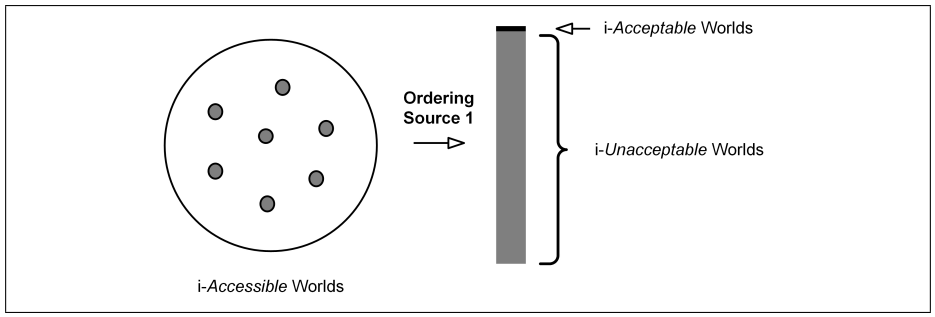


Figure 31: From ordered accessible worlds to the acceptable ones

top that way. So something else. For sake of concreteness: assume a ranking by *social welfare*. OK, so we are now back where we started, with the i-acceptable worlds ranked, but with a bit of light shed in the process I think. For now we are in a better position to ask:

> How then do we rank the missing *i*-unacceptables: a) by social welfare ranking or b) by justice ranking instead?

Classic contrary to duty (CTD) style reasoning serves here: Let X be the proposition that characterizes the de facto *i*-acceptable worlds. We ask what *would be* *i*-acceptable *given that* X is foreclosed (some *i*-unacceptable world will be accessed by our agent). It would seem that we must turn to the *next justice-best* worlds (See Figure 32).

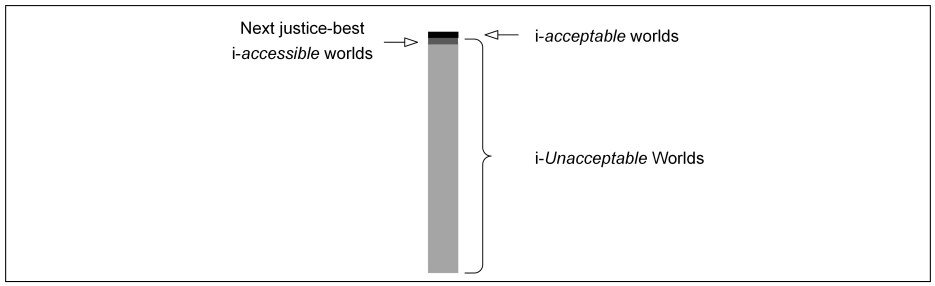


Figure 32: best and next-best justice worlds

This suggests a social welfare ranking of the *i*-unacceptables can't be right. For that might rank some neither best nor next-justice-best worlds above the next best justice ranked worlds, but that does not fit the implicit priority of the *first* ordering source over the *second*, say, the justice ranking over the social welfare ranking. So ranking the *i*-unacceptables as suggested by option a) above fails. That leaves option b): order them

by justice ranking. But b) can't be right either. For this would ignore morally relevant distinctions in *social* welfare ranking among worlds that are *equi-ranked* as *next-justice-best* worlds.

This is where CTD-style reasoning blends with that called for by supererogation and kin. The picture that emerges is displayed in Figure 33.

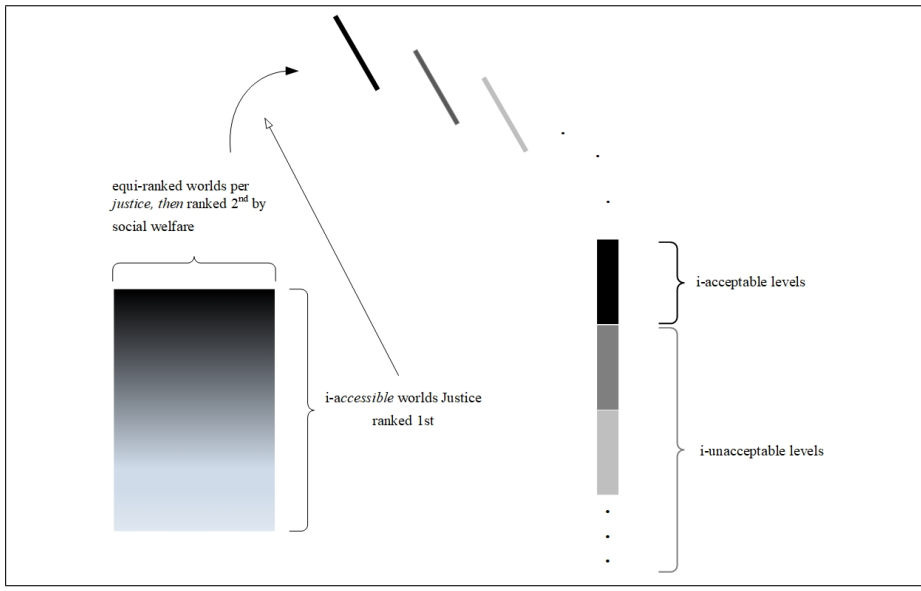


Figure 33: The generated composite ordering

Let us make this a bit more formal. Imagine we have three ordering relations, ${}^{1}\succsim_i$, ${}^{2}\succsim_i$, ${}^{3}\succsim_i$.

${}^{1}\succsim_i$: 1st ordering per i (e.g. justice)
${}^{2}\succsim_i$: 2nd ordering per i (e.g. social welfare)
${}^{3}\succsim_i$: 3rd composite ordering per i.

The first two are primitive (for now at least), but the third is not:

$$j \; {}^{3}\succsim_i k \stackrel{\text{def}}{=} j \; {}^{1}\succsim_i k \text{ or both } j \; {}^{1}\approx_i k \; \& \; j \; {}^{2}\succsim_i k.$$

where for each ordering relation $n(1, 2, 3)$, we assume these familiar definitions:

$$j \; {}^{n}\succ_i k \stackrel{\text{def}}{=} j \; {}^{n}\succsim_i k \; \& \; \neg(k \; {}^{n}\succsim_i j) \text{ and } j \; {}^{n}\approx_i k \stackrel{\text{def}}{=} j \; {}^{n}\succsim_i k \; \& \; k \; {}^{n}\succsim_i j.$$

Assume ${}^{1}\succsim_I$ and ${}^{2}\succsim_i$ are quasi-orders (reflexive and transitive relations) on the i-accessible worlds. These follow:

A) $j\ {}^{1}\succsim_i k \rightarrow j\ {}^{3}\succsim_i k$
B) $j\ {}^{3}\succsim_i k \rightarrow j\ {}^{1}\succsim_i k$
C) $j\ {}^{3}\succsim_i k \nrightarrow j\ {}^{2}\succsim_i k$
D) $j\ {}^{2}\succsim_i k\ \&\ j\ {}^{1}\approx_i k \rightarrow j\ {}^{3}\succsim_i k$
E) ${}^{3}\succsim_i$ is reflexive, and transitive: (i) $j\ {}^{3}\succsim_i j$ and (ii) $(j\ {}^{3}\succsim_i k\ \&\ k\ {}^{3}\succsim_i l) \rightarrow j\ {}^{3}\succsim_i l$
F) If ${}^{1}\succsim_i$ and ${}^{2}\succsim_i$ are also total, then ${}^{3}\succsim_i$ is total: $j\ {}^{3}\succsim_i k \vee k\ {}^{3}\succsim_i j$.

Returning to indifference, something will *now* be a matter of indifference iff it and its negation occurs somewhere within each ${}^{3}\succsim_i$*-based* i-level, that is, all i-levels, so *not*, simply among the i-acceptable levels. Anything that is a matter of indifference this way for the final ranking will be a) indifferent for the justice rankings, but not vice versa, and b) indifferent for each social welfare ranking of any given justice level, but not necessarily if we look at the social welfare ranking alone of all i-accessible worlds. These points reflect the priority.

We now have a fuller picture of how the structures motivated in the multipoint cumulative case argument might arise in an extension of those used in the more limited formal frames of DWE and DWE$^{\mathrm{m}}$. Working out the formalities here is left for a future occasion. We must also set aside an exploration of the fact that the formal DWE and DWE$^{\mathrm{m}}$ framework can be extended to cover generalizations of the operators. Not only can contrary to *duty* conditionals be modelled, but also contrary to *optimality* conditionals (if you are not going to give the colleague the delicate news in person, which you ought to but don't have to do, then you ought to give it by phone); likewise for contrary to *minimality* conditionals, etc. (See [McNamara, Forthcoming] for a first instalment.)

We turn next to an important extension of the DWE framework in another direction.

5.9 Aretaic (agent-evaluative) notions and DWE

What of supererogation and offences (suberogation)

Consider the following complaint about DWE: There is no representation (implicit or explicit) of aretaic notions (agent-evaluative notions) in DWE. Yet supererogation analytically entails *praiseworthiness*, a paradigmatically agent-evaluative concept. So the operator **B**C can't

represent supererogation. Also, if this typically endorsed equivalence is sound, ϕ *is supererogatory* iff ϕ is *beyond the call*, then DWE fails to represent either notion. Furthermore, what of the notion of an *offence* or *suberogation*? An offence entails *blameworthiness*, but once again, the latter is not expressible in the DWE framework, so neither can offences, the purported mirror image of supererogation, be expressed in that framework. Without agent appraisal, these concepts are not expressible in DWE. Although the claim that supererogation analytically entails praiseworthiness and that the equivalence above holds are not so straightforwardly obvious as they might seem (see [McNamara, 2011b]), the importance of extending the DWE framework to account for praiseworthiness and blameworthiness and derivative agent-evaluative notions is clear. We sketch next the picture outlined in [McNamara, 2000], and especially [McNamara, 2011a].

A simple preliminary framework for aretaic appraisal

We evaluate agents for actions, motives, traits, states of affairs, etc. Assume propositions can serve:

that Jane Doe performs action A / has motive M / intends I / ...

The basic idea will be that some propositions *reflect favorably* on people, others *unfavorably*, some *more favorably* than others, and some *neutrally*. Let's stick to *all things considered* appraisal of Jane Doe for propositions, but confined to propositions *consistent with Jane Doe's abilities at* i: CO_i. We imagine CO_i is derived from a standard accessibility relation, CO_{ij}, on worlds, W, read as j *is consistent with Doe's abilities in* i (which is not as strong as what is *within* Jane Doe's abilities — that would be what is consistent with her abilities *to bring about*). Now impose a *quasi-ordering* on CO_i, the propositions consistent with her abilities at i. Then for each pair of propositions, X and Y, in CO_i,

$X \geq_i Y$ iff X *reflects at least as well on Jane Doe as* Y (per i)

We introduce a corresponding operator, p $\geq$ q. *Strong preference* and *equi-ranking* relations are easily definable at both levels:

$$p > q \stackrel{\text{def}}{=} p \geq q \;\&\; \neg(q \geq p); \qquad p \approx q \stackrel{\text{def}}{=} p \geq q \;\&\geq p;$$
$$X >_i Y \stackrel{\text{def}}{=} X \geq_i Y \;\&\; \neg(Y \geq_i X); \quad X \approx_i Y \stackrel{\text{def}}{=} X \geq_i Y \;\&\; Y \geq_i X.$$

More familiar aretaic notions can then be defined as follows. We take tautological propositions to reflect neutrally on all agents. This is an anchor in the frames. We can then define these four notions:

> *Aretaically Neutral* (**A**N) propositions as those ranked equal to a tautology (for Jane Doe);
> *Aretaically indifferent* (**AI**) propositions as neutral ones with neutral negations;
> *Praiseworthy* (**P**W) propositions as ones ranked higher than a tautology;
> *Blameworthy* (**B**W) propositions as ones ranked lower than a tautology.

Figure 34 displays the intended modeling.

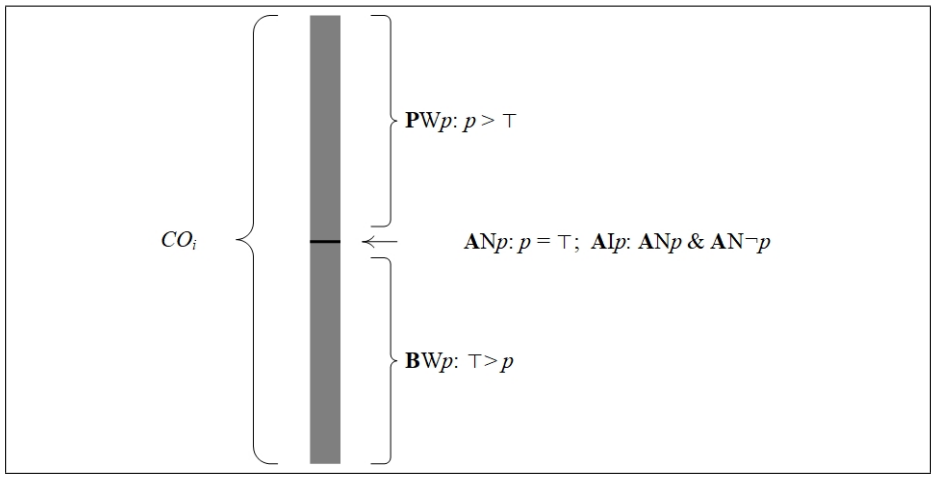


Figure 34: Semantics structure for **P**W, **B**W, **A**N, and **AI**

We gather all the propositions consistent with Jane Doe's abilities together and then rank them according to how well they reflect on Jane. Those that reflect better on Jane than tautologies, reflect favorably on her, and so she is to be evaluated as worthy of at least some degree of praise were such a proposition true; those ranked below a tautology reflect unfavorably on her, and so she is worthy of some degree of blame were such a proposition true; Those ranked equally with a tautology are aretaically neutrally for Jane (neither praise nor blame is associated with the proposition for Jane), and those whose negations also are aretaically neutral for Jane are aretaically indifferent for Jane.

We do not endorse $\geq$ -*Connectivity*,

$$\forall i \forall X \forall Y [X, Y \in CO_i \rightarrow (X \geq_i Y \vee Y \geq_i X)],$$

as basic (so the diagram above simplifies things). However, connectivity generates things presupposed in the classical framework for supererogation and kin, so let's assume it henceforth. Here are some consequences:

(CO-COMP) $(\mathrm{CO}p \,\&\, \mathrm{CO}q) \rightarrow (p \geq q \vee q \geq p)$
(AR-EXH) $\mathrm{CO}p \rightarrow (\mathbf{A}\mathrm{N}p \vee \mathbf{P}\mathrm{W}p \vee \mathbf{B}\mathrm{W}p)$
(AI-DEF′) $\mathbf{A}\mathrm{I}p \leftrightarrow (\mathrm{CO}p \,\&\, \neg\mathbf{B}\mathrm{W}p \,\&\, \neg\mathbf{B}\mathrm{W}\neg p \,\&\, \neg\mathbf{P}\mathrm{W}p \,\&\, \neg\mathbf{P}\mathrm{W}\neg p)$

Do all-in-all praiseworthiness and blameworthiness satisfy *PW-BW No Conflicts* principles?

(PW-NC) $\neg(\mathbf{P}\mathrm{W}p \,\&\, \mathbf{P}\mathrm{W}\neg p)$ (BW-NC) $\neg(\mathbf{B}\mathrm{W}p \,\&\, \mathbf{B}\mathrm{W}\neg p)$?

These are at least contenders for all-out aretaic appraisal and are presupposed in the classical conceptions of supererogation & offense, so let's assume they hold too. We can generate them by adding two constraints:

(PW-NC′) $\forall i \forall X (X >_i \top \rightarrow \neg(W - X >_i \top))$
(BW-NC′) $\forall i \forall X (\top >_i X \rightarrow \neg(\top >_i W - X))$

Simple aretaic partitions of CO_i for **P**W and for **B**W emerge (Figures 35 and 36).

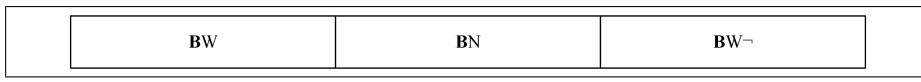


Figure 35: **B**W-based partition

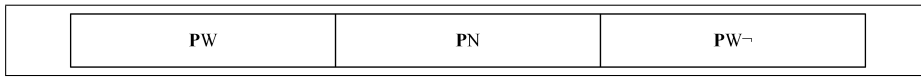


Figure 36: **P**W-based partition

where

$\mathbf{P}\mathrm{N}p \stackrel{\text{def}}{=} \neg\mathbf{P}\mathrm{W}p \,\&\, \neg\mathbf{P}\mathrm{W}\neg p$ (It is *Praise Neutral* that p)
$\mathbf{B}\mathrm{N}p \stackrel{\text{def}}{=} \neg\mathbf{B}W p \,\&\, \neg\mathbf{B}W\neg p$ (It is *Blame-Neutral* that p)

If we combine those two partitions, we get this sevenfold PW-BW aretaic partition (Figure 37).

	BW:	**B**N:	**B**W¬:
PW:	**P**W & **B**W	**P**W & **B**N	**P**W & **B**W¬
PN:	**P**N & **B**W	**A**I	**P**N & **B**W¬
PW¬:	**P**W¬ & **B**W	**P**W¬ & **B**N	**P**W¬ & **B**W¬

Figure 37: seven-fold aretaic partition

The two corner shaded cells are excluded (e.g. if p were ranked higher and lower than a tautology, by transitivity p would be ranked higher than p).

Adding a Simple Deontic Module

What happens if we blend in the standard threefold deontic partition below (Figure 38)?

MU	**O**P	**I**M

Figure 38: Traditional three-fold deontic classification

We get this 21-fold Aretaic-Deontic partition (Figure 39).

	PW & **B**N:	**P**W & **B**W¬:	**P**O & **B**W:	**A**I:	**P**O & **B**W¬:	**P**W¬ & **B**W:	**P**W¬ & **B**N:
MU:	**P**W & **B**N & **M**U	**P**W & **B**W¬ & **M**U	**P**O & **B**W & **M**U [elim by b)]	**A**I & **M**U	**P**O & **B**W¬ & **M**U	**P**W¬ & **B**W & **M**U [elim by a/b)]	**P**W¬ & **B**N & **M**U [elim by a)]
OP:	**P**W & **B**N & **O**P (**S**U^{a}) [elim by c)]	**P**W & **B**W¬ & **O**P (**Q**S) [elim by c)/d)]	**P**O & **B**W & **O**P (**O**F^{a}) [elim by d)]	**A**I & **O**P (**F**I)	**P**O & **B**W¬ & **O**P (**O**F$^{a\neg}$) [elim by d)]	**P**W¬ & **B**W & **O**P (**Q**O) [elim by c)/d)]	**P**W¬ & **B**N & **O**P (**S**U$^{a\neg}$) [elim by c)]
IM:	**P**W & **B**N & **I**M [elim by a)]	**P**W & **B**W¬ & **I**M [elim by a)/b)]	**P**O & **B**W & **I**M	**A**I & **I**M	**P**O & **B**W¬ & **I**M [elim by b)]	**P**W¬ & **B**W & **I**M	**P**W¬ & **B**N & **I**M

Figure 39: twenty-one-fold aretaic-deontic partition

With an aretaic and deontic module, we can explore the logic of a variety of moral conditions of interest via this simple framework. A general question:

How are deontic and aretaic conditions related, and how far can deontic and aretaic valences diverge or how closely must they match?

Consider these bridging principles:

a) No **P**W-**I**M Conflicts: $\neg(\mathbf{P}\mathrm{W}p \,\&\, \mathbf{I}\mathrm{M}p)$ [i.e. $\mathbf{P}\mathrm{W}p \rightarrow \mathbf{P}\mathrm{E}p$]
b) No **B**W-**O**B Conflicts: $\neg(\mathbf{B}\mathrm{W}p \,\&\, \mathbf{O}\mathrm{B}p)$ [i.e. $\mathbf{B}\mathrm{W}p \rightarrow \mathbf{P}\mathrm{E}\neg p$]

These are reductive. The six eliminations entailed by each are among the shaded boxes in the top and bottom rows: a) eliminates the last two shaded cells of the top row, and the first two of the bottom row; b) eliminates the first two shaded cells of the top row and the last two shaded cells of the bottom row. The result of adding both principles is a partition with only 15 deontic-aretaic positions (those lightly shaded as well as those unshaded).

The standard account of supererogation and offence and Mellema's extensions

The *classical analyses* of supererogation and offense are easy to define:

$\mathbf{S}\mathrm{U}^{\mathrm{a}}p \stackrel{\mathrm{def}}{=} \mathbf{O}\mathrm{P}p \,\&\, \mathbf{P}\mathrm{W}p \,\&\, \neg\mathbf{B}\mathrm{W}\neg p$
$\mathbf{O}\mathrm{F}^{\mathrm{a}}p \stackrel{\mathrm{def}}{=} \mathbf{O}\mathrm{P}p \,\&\, \mathbf{B}\mathrm{W}p \,\&\, \neg\mathbf{P}\mathrm{W}\neg p$

Mellema [1987, 1991] argues for acts of *quasi-supererogation* and *quasi-offense*:

$\mathbf{Q}\mathrm{S}p \stackrel{\mathrm{def}}{=} \mathbf{O}\mathrm{P}p \,\&\, \mathbf{P}\mathrm{W}p \,\&\, \mathbf{B}\mathrm{W}\neg p$
$\mathbf{Q}\mathrm{O}p \stackrel{\mathrm{def}}{=} \mathbf{O}\mathrm{P}p \,\&\, \mathbf{B}\mathrm{W}p \,\&\, \mathbf{P}\mathrm{W}\neg p$

Finally, we introduce one more mixed concept, (deontically) *optional aretaic indifference*:

$\mathbf{O}\mathrm{I}p \stackrel{\mathrm{def}}{=} \mathbf{O}\mathrm{P}p \,\&\, \mathbf{A}\mathrm{I}p$

We could also sensibly introduce *weak-supererogation* and *weak-offense*:

$\mathbf{W}\mathrm{S}p \stackrel{\mathrm{def}}{=} \mathbf{O}\mathrm{P}p \,\&\, \mathbf{P}\mathrm{W}p$ (i.e. $\mathbf{S}\mathrm{U}p \vee \mathbf{Q}\mathrm{S}p$)
$\mathbf{W}\mathrm{O}p \stackrel{\mathrm{def}}{=} \mathbf{O}\mathrm{P}p \,\&\, \mathbf{B}\mathrm{W}p$ (i.e. $\mathbf{O}\mathrm{F}p \vee \mathbf{Q}\mathrm{O}p$)

(The classical style analysis of **S**U and **O**F above can be shown to fail. See [Hansson, 2001] (and earlier references there) and [McNamara, 2011b] for arguments.)

Normative positions in these frameworks and some reductive schemes

The five new operators defined above are already in our prior twenty-one-fold partition (see middle row, parentheticals). So lingering behind the classical conception of supererogation is a potential for at least 21 mutually exclusive, jointly exclusive categories (normative positions)–far more than previously articulated.

Now, some argue for the rejection of supererogation by endorsing:

c) $\mathbf{P}\mathrm{W}p \rightarrow \mathbf{O}\mathrm{B}p$

c) entails $\neg(\mathbf{S}\mathrm{U}^{\mathrm{a}}p \vee \mathbf{Q}\mathrm{S}p \vee \mathbf{Q}\mathrm{O}p)$, thus eliminating the first two and the last two of the lightly shaded cells in the middle row. Only the three middle cells of the middle row would remain (only offences and the fully indifferent categories in the middle would remain. Given *No Conflicts* for **O**B, c) also entails our earlier a) $\mathbf{P}\mathrm{W}p \rightarrow \mathbf{P}\mathrm{E}p$, so the last two shaded cells of the top row, and the first two of the bottom row go too. Principle c) is thus *highly* eliminative.

The following bridging principle is often endorsed (in arguing for the rejection of suberogation, but also in recent discussions of determinism):

d) $\mathbf{B}\mathrm{W}p \rightarrow \mathbf{I}\mathrm{M}p$

d) entails $\neg(\mathbf{O}\mathrm{F}^{\mathrm{a}}\ p \vee \mathbf{Q}\mathrm{S}p \vee \mathbf{Q}\mathrm{O}p)$, thus eliminating all but the central and end cells of the middle row. Only the supererogation and full indifference categories remain in the middle row. Given *No Conflicts* for **O**B, d) also entails our earlier b) $\mathbf{B}\mathrm{W}p \rightarrow \mathbf{P}\mathrm{E}\neg p$, so the first two shaded cells of the top row and the last two shaded cells of the bottom row also go. So principle d) is also highly eliminative.

In sum, if c) or d) hold, Mellema's quasi-notions are out, along with supererogation or suberogation, and two of the three shaded boxes in the top and bottom rows; in each case, there are eight eliminations, leaving only thirteen positions; if both c) and d) hold, there are twelve eliminations, leaving just the nine white unshaded positions.

Integrating with DWE and expanding the normative positions

What happens if we combine the twenty one aretaic normative positions above with DWE's deontic positions? If we make the combination in question, we have seven aretaic positions and twelve DWE positions, so we get eight-four combined positions provided we add no eliminating additional claims like those above (Figure 40).

	PW & BN:	PW & BW¬:	PN & BW:	AI:	PN & BW¬:	PW¬ & BW:	PW¬ & BN:
MU:	PW & BN & MU	PW & BW¬ & MU	PN & BW & MU	AI & MU	PN & BW¬ & MU	PW¬ & BW & MU	PW¬ & BN & MU
OU & LE¬:	PW & BN & OU & LE¬	PW & BW¬ & OU & LE¬	PN & BW & OU & LE¬	AI & OU & LE¬	PN & BW¬ & OU & LE¬	PW¬ & BW & OU & LE¬	PW¬ & BN & OU & LE¬
OU¬ & LE¬:	PW & BN & OU¬ & LE¬	PW & BW¬ & OU¬ & LE¬	PN & BW & OU¬ & LE¬	AI & OU¬ & LE¬	PN & BW¬ & OU¬ & LE¬	PW¬ & BW & OU¬ & LE¬	PW¬ & BN & OU¬ & LE¬
¬OU & ¬OU¬ & LE¬:	PW & BN & ¬OU & ¬OU¬ & LE¬	PW & BW¬ & ¬OU & ¬OU¬ & LE¬	PN & BW & ¬OU & ¬OU¬ & LE¬	AI & ¬OU & ¬OU¬ & LE¬	PN & BW¬ & ¬OU & ¬OU¬ & LE¬	PW¬ & BW & ¬OU & ¬OU¬ & LE¬	PW¬ & BN & ¬OU & ¬OU¬ & LE¬
OU & LE:	PW & BN & OU & LE	PW & BW¬ & OU & LE	PN & BW & OU & LE	AI & OU & LE	PN & BW¬ & OU & LE	PW¬ & BW & OU & LE	PW¬ & BN & OU & LE
OU¬ & LE:	PW & BN & OU¬ & LE	PW & BW¬ & OU¬ & LE	PN & BW & OU¬ & LE	AI & OU¬ & LE	PN & BW¬ & OU¬ & LE	PW¬ & BW & OU¬ & LE	PW¬ & BN & OU¬ & LE
¬OU & ¬OU¬ & LE:	PW & BN & ¬OU & ¬OU¬ & LE	PW & BW¬ & ¬OU & ¬OU¬ & LE	PN & BW & ¬OU & ¬OU¬ & LE	AI & ¬OU & ¬OU¬ & LE	PN & BW¬ & ¬OU & ¬OU¬ & LE	PW¬ & BW & ¬OU & ¬OU¬ & LE	PW¬ & BN & ¬OU & ¬OU¬ & LE
OU & ¬LE & ¬LE¬:	PW & BN & OU & ¬LE & ¬LE¬	PW & BW¬ & OU & ¬LE & ¬LE¬	PN & BW & OU & ¬LE & ¬LE¬	AI & OU & ¬LE & ¬LE¬	PN & BW¬ & OU & ¬LE & ¬LE¬	PW¬ & BW & OU & ¬LE & ¬LE¬	PW¬ & BN & OU & ¬LE & ¬LE¬
OU¬ & ¬LE & ¬LE¬:	PW & BN & OU¬ & ¬LE & ¬LE¬	PW & BW¬ & OU¬ & ¬LE & ¬LE¬	PN & BW & OU¬ & ¬LE & ¬LE¬	AI & OU¬ & ¬LE & ¬LE¬	PN & BW¬ & OU¬ & ¬LE & ¬LE¬	PW¬ & BW & OU¬ & ¬LE & ¬LE¬	PW¬ & BN & OU¬ & ¬LE & ¬LE¬
¬OU & ¬OU¬ & ¬LE & ¬LE¬:	PW & BN & ¬OU & ¬OU¬ & ¬LE & ¬LE¬	PW & BW¬ & ¬OU & ¬OU¬ & ¬LE & ¬LE¬	PN & BW & ¬OU & ¬OU¬ & ¬LE & ¬LE¬	AI & ¬OU & ¬OU¬ & ¬LE & ¬LE¬	PN & BW¬ & ¬OU & ¬OU¬ & ¬LE & ¬LE¬	PW¬ & BW & ¬OU & ¬OU¬ & ¬LE & ¬LE¬	PW¬ & BN & ¬OU & ¬OU¬ & ¬LE & ¬LE¬
IN:	PW & BN & IN	PW & BW¬ & IN	PN & BW & IN	AI & IN	PN & BW¬ & IN	PW¬ & BW & IN	PW¬ & BN & IN
IM:	PW & BN & IM	PW & BW¬ & IM	PN & BW & IM	AI & IM	PN & BW¬ & IM	PW¬ & BW & IM	PW¬ & BN & IM

Figure 40: DWE-aretaic eight-four-fold partition

This includes improved conditions for the classical analysis of supererogation and suberogation (offence):

SUp: **B**Cp & **P**Wp & ¬**B**W¬p
OFp: **P**Sp & **B**Wp & ¬**P**W¬p

This in turn allows for a distinction between going beyond the call and

supererogation (as adding an *agent*-evaluative element to doing more good than you have to), and similarly for suboptimality and suberogation/offence [McNamara, 2011b].

We turn next to making a quick assessment of the contributions sketched in Section 5.

5.10 Reflections on the DWE framework

The DWE framework is the first model-theoretic framework that, to a first approximation, gives a reasonable representation of many key features of a cluster of core concepts of common morality that reside in the neighborhood of supererogation. It distinguishes pairs of concepts that have often been conflated with one another, what is a "must" versus an "ought", what is optional versus indifferent, it also brings attention to the neglected notion of "the least one can do", it provides representations of not only what is beyond the call, but also what is permissibly suboptimal, and it provides a substantial enrichment of the classical deontic scheme of concepts, as evidenced for example in the substantially expanded normative positions (as illustrated by the DWE's twelvefold classification diagram). Nonetheless, there are a variety of places where revisions are called for and expansions would add value (or already exist but must be set aside here).

1) The DWE framework is classical. Even DWE^{m} retains some key classical features: although it allows for conflicting *oughts* and *leasts*, and does not endorse aggregation for either, it nonetheless ratifies necessitation for *must*, ought, and *least*, and more troublingly for many, it ratifies RM (Inheritance) for the three operators. RM is often blamed for various paradoxes in the case of obligations.[91]

2) More specifically, DWE and DWE^{m} are subject to a *Paradox of Disjunctive Supererogation* to be discussed more below.

3) There is no *explicit axiological* notion/s (e.g. good, bad, better) represented in the object language, yet the intuition for the analysis of supererogation is that of "doing more good than one would have done had one done the least one could have done". At best, only the model theory represents any axiological notion, namely

[91] For example, RM is often blamed for Ross' Paradox and for the Good Samaritan Paradox. See [Hilpinen and McNamara, 2013] from the first volume of this handbook.

the comparative i-relative world ordering relation, but none explicitly in the object language. One possibility is to recast the aretaic framework so that the basic primitive relation is axiological and then try to integrate that with the DWE framework, but that and other options must await another occasion.

4) There is no representation of agency or action. This is both a limit, and furthermore it raises questions about whether or not the operators can represent their intended targets concepts, and how to read the operators (or what they can be taken to represent). Section 5.7 attempted to address the latter, but not the former issue.

5) As we saw, the initial informal presentation of the semantics for indifference in Section 5.1 ranged over all (i-relative) levels of worlds, including unacceptable levels, yet in the formal semantics above, indifference is characterized only by reference to the acceptable worlds, with no reference to what is going on in the unacceptable ones.

6) There is no representation of conditional versions of the system (save briefly in McNamara's dissertation, and in some presentations), which would generalize the notion of contrary to duty conditionals to contrary to normative status conditionals (but see [McNamara, Forthcoming]).

The Paradox of Disjunctive Supererogation, is worth a brief separate discussion. Recall that DWE defines BC as follows:

$$\mathbf{B}\mathrm{C}\phi \stackrel{\text{def}}{=} \mathbf{L}\mathrm{E}\neg\phi \;\&\; \mathbf{P}\mathrm{E}\phi$$

It is easy to prove in DWE and $\mathrm{DWE}^{\mathrm{m}}$ the following theorem saying that if ψ is impermissible but ϕ is beyond the call, then their disjunction is beyond the call:

$$\vdash \mathbf{I}\mathrm{M}\psi \;\&\; \mathbf{L}\mathrm{E}\neg\phi \;\&\; \mathbf{P}\mathrm{E}\phi. \rightarrow \mathbf{B}\mathrm{C}(\phi \vee \psi)$$

For the antecedent's third conjunct entails $\mathbf{P}\mathrm{E}(\phi \vee \psi)$, and the first and second together entail $\mathbf{L}\mathrm{E}\neg(\phi \vee \psi)$, so the defining condition for $\mathbf{B}\mathrm{C}$ is met not just by ϕ, but by the disjunction $\phi \vee \psi$. In our stock rescue example, the least I can do is pull the fire alarm and that precludes *rescuing Tiny Tim* (t), which we assume is permissible and genuinely

beyond the call. But then assuming *fanning the flames* (f) is impermissible, it follows that it is beyond the call for Doe to be such that either she rescues Tiny Tim or fans the flames, $\mathbf{B}\mathrm{C}(\mathrm{t} \vee \mathrm{f})$. But $\mathrm{t} \vee \mathrm{f}$ can be as easily realized via f as via t, and thus as easily realized via the impermissible as the permissible. So we get an analogue to the Ross paradox–one can supererogate by fanning the flames.

I have sketched possible solutions to this in conference presentations, but it is beyond the scope of this essay to explore this in any detail other than to make two quick notes in passing. One is that if we dropped weakening (RM) for **O**B, then $\mathbf{P}\mathrm{E}\phi$ no longer yields $\mathbf{P}\mathrm{E}(\phi \vee \psi)$, so the paradox is blocked. There are ways to do this while preserving much of the flavor and attractions of the DWE framework. Let me also note that the addition of an evaluative operator, in line with one of Hansson's contentions, could block it as well. A monadic operator for *it would be good that* (**G**D) might suffice. For if we added that condition to the defining condition above, then as long as **G**D is not itself subject to RM, the paradox will be blocked. However, one way to see the paradox is that it indicates that the DWE analysis of going beyond the call does not capture the part of the informal gloss of doing more good than one would if one did the least, for this is just what is not assured by acting so that either the child is rescued or the flames are fanned.[92] This however invites further questions about how maximally specific must one be in order to say something is or would be beyond the call. We must leave these important issues for elsewhere.

Before leaving this Section, let me note that the DWE framework has been recast in an Andersonian-Kangerian framework (see [Hilpinen and McNamara, 2013, Sections 6.2 and 7.2] for background), with associated determination theorems, as well as assessment of some limitations [McNamara, 1999]. Humberstone [Humberstone, 1974] sketched an early variation of Anderson's reduction where he used two constants rather than one to provide more space for some of the concepts we have focused on. Lastly, [Forrester, 1975] outlines a framework to also make more space for some of the concepts we have been exploring, one that can be fruitfully recast via Andersonian-Kangerian concepts. We regret that space limitations preclude what might otherwise be the natural inclusion and exploration of this work here.

[92]One possibility is to reinterpret the aretaic framework in Section 5.9, so that the basic primitive relation is axiological instead, and then, with other adjustments, try to integrate that in a framework with DWE's flavor.

6 Some other recent work

6.1 Wessels' work, and supererogatory holes

Criticism of standard threshold model via supererogatory holes

Ulla Wessels [Wessels, 2002; Wessels, 2003; Wessels, 2015] introduces a novel puzzle case, an argument based on that case against what she calls "The Threshold Model", and a correlative thesis for the existence (in principle) of what she calls "Supererogatory Holes". She then turns to the question of what the structure of one's moral options must be if supererogatory holes are to be accounted for, and proposes what she calls "The Format", a hypothesis about the structure called for in an account of supererogation that allows for supererogatory holes. Her work is, in my estimation, significant and undeservedly not better known, and so I am happy to exposit some central aspects of it here. I will rely primarily on [Wessels, 2015] and on associated unpublished materials that Wessels kindly provided.

The Threshold Model (TM) is defined as follows: in every situation there is a threshold for the good to be done such that 1) it is obligatory to perform an action that meets the threshold (is at least that good) and 2) every action that exceeds the threshold (i.e. does more good) is supererogatory [Wessels, 2015, p. 88].[93] Wessels then asks us to consider the following case:

possible actions:	donations in €:	numbers of lives saved by the actions:
A_5	10,050	200
A_4	10,000	101
A_3	5,000	100
A_2	50	1
A_1	0	0

For sake of argument, we imagine that the first option is impermissible, since we are obligated to meet or exceed the threshold–donating at least 50€, which will save one life, but we are not required to do more than

[93] This does not quite fit the nuances of positions with a flavor like Slote's (Cf. [Hurka, 1990]), where we might say that there is a minimal threshold of good, *which if achievable*, we must achieve, but need not go beyond, but if not achievable (if being below the threshold is inevitable) in some situation, then we are obliged to optimize.. However, I think this does not affect the spirit of the main point to be made which will apply in all cases where there is a threshold and could be so adapted to apply to Slote.

that, so that A_3 for example is beyond the call. The key here is to focus on the relationship between the values associated with A_4 and A_5. We are to imagine that the agent is able to donate any of the amounts listed in the middle column. Furthermore, although the second highest donation of 10000€ would be costly to the agent (though not approaching catastrophic), the difference in cost to the agent between that donation and A_5 (adding just 50€ more — the minimal amount) would result in only a marginal decrement in agent-utility. Wessels then contends that although A2, A3, and A5 are permissible, and the latter two supererogatory, A4, though it saves more lives than A_2 and A_3, is *not* supererogatory because it is *not even permissible* [Wessels, 2015, p. 90]. Roughly, if you go so far as A_4, then it is unacceptable to not take the remaining small step to A_5 and thereby save 99 more lives for just 50€ more. Thus the case, she suggests, reveals a fundamental flaw for TM, since there can always be cases like that above, where not all options above the threshold are supererogatory since not all of them are permissible, and so TM's second clause is violated.

Wessels considers various objections to her argument against TM (both in [Wessels, 2015], and more objections in [Wessels, 2002], but for our purposes, we pass over these, other than to draw out the key argument for the conclusion that A_4 is impermissible and hence not supererogatory, an interesting argument derived from a conditional obligation claim about the case along with some plausible principles of conditional deontic logic. To make the structure of the argument clear, let us introduce the following abbreviations:

T: The agent donates *exactly* 10,000€ (i.e. she performs A_4).
T′: The agent donates *at least* 10,000€
T″: The agent donates *at least* 10,050€

The key conditional obligation claim is that it is obligatory that the agent gives at least 10050€ if she gives at least 10000€, that is, $\mathbf{O}\mathrm{B}_{\mathrm{T}'}\mathrm{T}''$. The argument is then the following, where "deontic principle" below is short for "deemed an instance of a general valid deontic principle".[94]

[94]The general valid principles in question are just those resulting from replacing the constants T, T′, and T″ with variables and viewing them as schemata.

1) $\mathbf{O}\mathrm{B}_{\mathrm{T}'}\mathrm{T}''$ [plausible premise about the case]
2) $\mathbf{O}\mathrm{B}_{\mathrm{T}'}\mathrm{T}'' \rightarrow \mathbf{O}\mathrm{B}\neg(\mathrm{T}' \,\&\, \neg\mathrm{T}'')$ [deontic principle][95]
3) So, $\mathbf{O}\mathrm{B}\neg(\mathrm{T}' \,\&\, \neg\mathrm{T}'')$
4) T entails $(\mathrm{T}' \,\&\, \neg\mathrm{T}'')$ [analytic premise]
5) If T entails $(\mathrm{T}' \,\&\, \neg\mathrm{T}'')$, then $\mathbf{O}\mathrm{B}\neg(\mathrm{T}' \,\&\, \neg\mathrm{T}'') \rightarrow \mathbf{O}\mathrm{B}\neg\mathrm{T}$ [deontic principle]
6) So, $\mathbf{O}\mathrm{B}\neg\mathrm{T}$ (That is, A_4 is impermissible).[96]

So assuming a threshold of 50€ for sake of argument in the above example, there is a *supererogatory hole*: an alternative ranked above a supererogatory alternative that is not itself a supererogatory alternative (despite being better–doing more good). This is inconsistent with TM clause 2). Let me point out two things here. First, ignoring the issue of supererogation, *at a more fundamental level*, this is an argument against the very widely endorsed principle that *anything better than something permissible is permissible*, and so has wider significance. Secondly, although Wessels is arguing against the *traditional* threshold account, this is not to say that she rejects the idea that a *non-traditional* account employing a notion of a threshold is called for; on the contrary, as we shall see.

The Format: Wessels' hole-allowing account of supererogation

The basic strategy for *The Format* (TF) is to identify structural features of ranked moral alternatives, and define a two place technical relation, *A is supererogatory relative to* A' (in a situation), and then use that to define when an alternative *A is supererogatory* per se (in a situation). TF is somewhat complicated, so I will sketch things out for the binary predicate and some of the notions to be employed; then I will turn to TF itself.

A central notion for Wessels is that there is something she calls the "*supererogatory measure*" (a real number) related to two alternatives, A and A' open to an agent. She imagines that this supererogatory measure (SM) is a real number function of four real number values: in the intended interpretation, the input numbers would measure the agent

[95] This is valid in most accounts of conditional obligations and seems quite intuitive. The principle behind 5) is RM, a contested, but often endorsed, principle. In either event, the argument is an interesting one, and she could perhaps add a material premise instead of 4) and 5) saying in this or such cases $\mathbf{O}\mathrm{B}\neg(\mathrm{T}' \,\&\, \neg\mathrm{T}'') \rightarrow \mathbf{O}\mathrm{B}\neg\mathrm{T}$ holds.

[96] Cf. Appendix A in [Wessels, 2002].

utility (au) of each of two alternatives, A and A', and the total utility (tu) of each of the two alternatives, and then output a composite value based on their relationship. Roughly, it measures the comparative cost/-gain (if any) in agent-utility of choosing A *over* A' and compares that in turn to the comparative gain (if any) in total utility of choosing A' and A, and it outputs a number.[97] The idea is that if the cost to the agent is low in choosing say A' over A (so that $au(A') - au(A)$ is small) while the gain in moral value in that choice is high (so that $tu(A') - tu(A)$ is large), then the supererogatory measure (SM) of $\langle tu(A'), tu(A), au(A), au(A')\rangle$ will be low or nil.[98] Contrast that to a case where the things are reversed and the gain in moral value is small but the cost to agent is high, then $SM(\langle tu(A'), tu(A), au(A), au(A')\rangle)$ will itself be high. We can think of the supererogatory measure as a measure of the tendency of the choice of the first alternative over the second to be supererogatory, or equivalently, to be required (degree of demandingness).

Certain general constraints are then placed on the function SM. If either the value $tu(A')$ or $au(A')$ increases, then the SM value decreases (for if the first, then the moral utility A' gains over that of A increases the "demand-pressure" on the agent, and if the second, then the agent utility going up for the morally better option means the cost to the agent of that choice is going down, so once again the "demand-pressure" goes up). Thus, SM monotonically decreases in the 1st and 4th arguments of the function. In contrast, SM monotonically increases in the 2nd and 3rd arguments of the function, for if the values of $tu(A)$ or of $au(A)$ increase, so does the value of SM, for in the first case, the moral gain A' provides over A shrinks so that the demand for sacrifice on an agent to choose A' shrinks, and thus the merit of doing so anyway increases, and in the second case, the cost to the agent of choosing A' over A increases. and thus the sacrifice is more meritorious. We need one further constraint on the structures. As noted, there will still be a threshold, z, a real number, and substantive theories are to establish what this might be and how it is arrived at. One intuition behind this analysis, one often (but not always) attributed to supererogatory acts, is that some sacrifice must be involved on the part of the agent. So

[97] I use "total utility" to aid processing via something familiar, but "moral utility" or "overall moral value" fit the level of generality Wessel intends.

[98] The arguments, A and A', of the two utility functions are reversed in the third and fourth position to reflect the fact that the main interest is in cases where morality pulls in one direction, and agent utility pulls in the other, so that A' is better for morality than A, whereas things are reversed for agent utility, A is better than A' for the agent. There is nothing essential here about the order formally it just stresses the opposing pulls in focus here.

that A' is to be supererogatory relative to A only if the agent utility favors choice A, that is, $au(A) > au(A')$. Similarly, for an action to be supererogatory, the overall value of A' must be at least as great as that of A, $tu(A') \geq tu(A)$. The author suggests that A' need not be greater than A by citing a situation much like one this author has used, indicating that there could be someone in the wings who stands ready to make the same sacrifice, yielding the same overall value, but the agent chooses to take on the burden herself.[99] These two constraints on the *SM* function will be (negatively) cast in terms of the threshold: if either constraint is *not* met, then the supererogatory measure of the ordered set of four associated values must turn out below the threshold. More exactly the structures must be such that if either $au(A) \leq au(A')$ or $tu(A') < tu(A)$, then $SM(\langle tu(A'), tu(A), au(A), au(A')\rangle)$ will be cast below the threshold. In these two cases, the intuition is that choosing A' or A is unacceptable.

Lastly, there is always a background finite set of actions (the set of relevant alternatives) that a given action will be supererogatory with respect to. Before stating the full account, here is a helpful diagram that Wessels' gives [Wessels, 2015, p. 99] that can be used for previewing the overall account of supererogation per se via *A is supererogatory relative to A′* (See Figure 41). It is roughly an idealized decision procedure.
So A_j is supererogatory iff 1) there is an A_i such that A_j is supererogatory relative to A_i, 2) all actions ranked better than A_j are supererogatory relative to A_j, and 3) likewise for A_i (all action ranked better than A_i are supererogatory relative to A_i). Let us return to the motivating example, and see how this might apply informally. First consider A_4, giving exactly 10000€. We are imagining that there is an action, A_2 (giving 50€), such that A_4 is supererogatory relative to it, and thus the answer to Q1 is affirmative. (In fact, in this case, it is stipulated that all actions ranked above A_2 consist in doing more good than required, recall, but that is now deemed consistent with being impermissible.) However, the answer to question Q2 we imagine is negative, since A_5 (giving 10050€ and saving 99 more lives for just 50€ more) is *not* supererogatory relative to A_4 (though ranked higher than A_4), since A_5 is

[99] "... what if there is now a second potential rescuer, and if the mailwoman didn't go in, he would–where the resulting outcomes would be in parity? In such a case, our mailwoman has optimal alternatives where she goes in and optimal alternatives where she doesn't. So going in is not required in order to optimize. (Indeed, even a *utilitarian* would have to deem her rescue optional!) Now suppose she places her hand on the man's shoulder, says "I'll go", and does. As a result, he stays behind. Her sacrifice was still surely supererogatory." [McNamara, 1996b, p. 433].

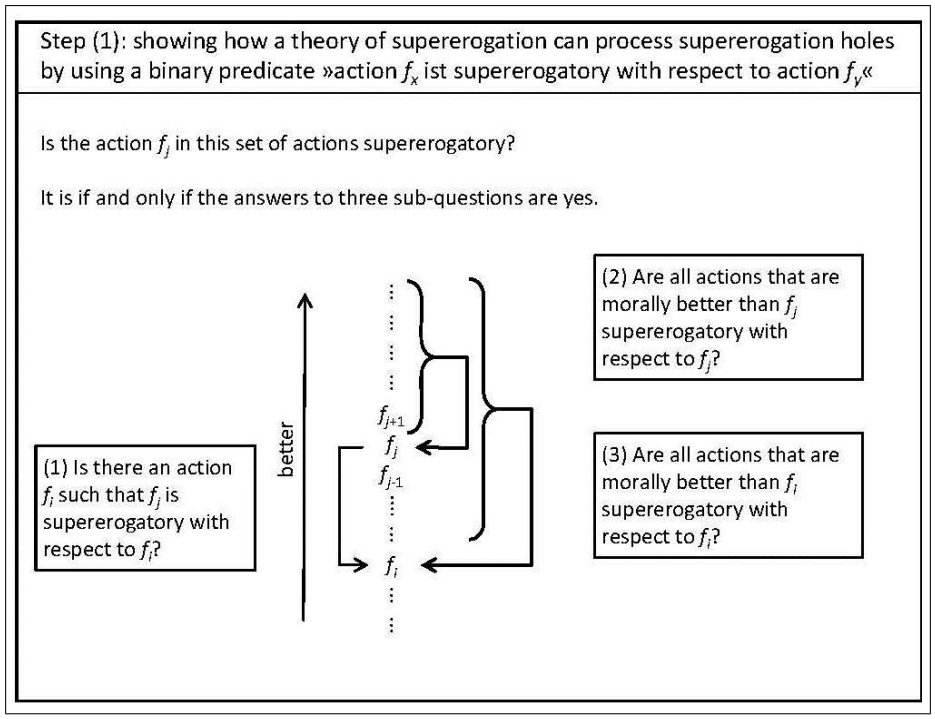


Figure 41: Wessell's account as 3 step procedure

obligatory relative to A_4 — the demand pressure is great between them, since the total utility gains are great, the agent loss is very small. Thus the above rules out A_4 as supererogatory per se. But what of the other actions deemed intuitively supererogatory in the case imagined? Consider A_5 itself. We are to imagine that it is supererogatory relative to A_2 and since it the highest ranked option, vacuously, Q2 are answered affirmatively. Regarding Q3, the idea is that all actions morally better than A_2, even non-supererogatory per se A_4, are supererogatory *relative* to A_2. So A_5 is supererogatory per se. Similarly, for A3. A question remains. What does Q3 add? Why is it essential? The thought is that we could be in a situation where A_i itself is like A_4 in our example, in that there is a much better alternative to A_i that places little additional burden on the agent, and then in that case, A_i is itself not permissible and so it can't anchor A_j as supererogatory per se. [Wessels, 2015, p. 101]

I now turn to formally stating TF:

$\exists$ a threshold z $\in R$
$\exists$ a function SM: $\mathrm{R}^4 \rightarrow R$
$\forall$ agent a
$\forall$ set of actions $A = \{A_1, \ldots, A_n\}_a$ available to a, and action A_j in that set

(F1) SM monotonically decreases in the 1st and 3rd argument
(F2) SM monotonically increases in the 2nd and 4th argument
(F3) $\forall\, x_1, x_2, x_3, x_4 \in R$: $SM(x_1, x_2, x_3, x_4) < z$, if $x_1 < x_2$ or $x_3 \geq x_4$
(F4) $\mathrm{super}_A(A_j)$ iff:
  (F4.1) a) $\exists A_i \in A$: $SM(\ tu(A_j),\ tu(A_i),\ au(A_i),\ au(A_j)) > z$ &
  b) $\forall A_k \in A$: $tu(A_k) > tu(A_i) \rightarrow SM(tu(A_k),\ tu(A_i),\ au(A_i),\ au(A_k)) > z$, and
  (F4.2) $\forall A_k \in A$: $tu(A_k) > tu(A_j) \rightarrow SM(tu(A_k),\ tu(A_j),\ au(A_j),\ au(A_k)) > z$.

So there is a threshold (z) and a function (SM) such that for any agent and set of alternatives, and any one of those alternatives A_j, the supererogatory measure function has the properties discussed earlier (F2)-(F3), and A_j is supererogatory with respect to the set of alternatives, A, iff the answer to the three questions discussed above is affirmative, which is what F4 expresses essentially.

Let's return to our prior examples where we went through how the three questions would be answered for A_3-A_5, but now let's do so in a more formal way with an eye on TF.[100] We will sketch what the models might look like, in part, where we ignore some features of the models (e.g. the exact value of z, the values the functions au and tu provide, which are left tacit by Wessels). Here I follow some material Wessels kindly provided. First consider A_4, giving exactly 10000€. It is to come out as *non*-supererogatory for the reasons mentioned earlier. All models with the following features will yield the desired results. $A = \{A_1, \ldots, A_5\}$ and $tu(A_5) > tu(A_4) > tu(A_3) > tu(A_2) > tu(A_1)$. We assume SM is made to satisfy the constraints stipulated in F1-F3, and we focus on F4. Let $SM(tu(A_4), tu(A_2), au(A_2), au(A_4)) > z$, so F4.1 a) is satisfied, A_4 is supererogatory relative to A_2. Let $SM(tu(A_5), tu(A_2), au(A_2), au(A_5)) > z$, $SM(tu(A_4), tu(A_2), au(A_2), au(A_4)) > z$, and $SM(tu(A_3),\ tu(A_2),\ au(A_2),\ au(A_3)) > z$, so F4.1 b) is satisfied–each action in A ranked

[100] Cf. Appendix B in [Wessels, 2002].

above A_2 is supererogatory relative to A_2. However, make F4.2 not satisfied, $SM(tu(A_5), tu(A_4), au(A_4), au(A_5)) \leq z$ (not *above* the threshold), so A_5 is *not* supererogatory relative to A_4. Thus, A_4 is *not* supererogatory simpliciter in any models with these features. Consider A_5 as intended in the case. It comes out as supererogatory simpliciter in the class of models above, since that $SM(tu(A_5), tu(A_2), au(A_i), au(A_5)) > z$ was stipulated in assuring F4.1 b) above, so F4.1 a) holds for A_5 as well as A_4 relative to A_2, and F4.1 b) carries over unchanged, since A_2 is still the anchor, and F4.2 holds vacuously now, since no alternative has higher utility than A_5. Similarly for A_3.

Note that if we contracted the alternative set A so the A_5 is removed, F4.1 a) would be unaffected, F4.1 b) would still be satisfied, but now F4.2 is vacuously satisfied. For this contracted alternative set, A_4 is not a supererogatory hole, but just plain supererogatory.

Space limitations prevent me from further exploring this account or trying to begin to develop what a logic for this account might look like, be it cast act-theoretically or recast in propositional terms. The context for the account is primarily philosophical, and a logical proof system is not formally specified by Wessels as generated by her account of supererogation. Of course other questions remain, such as the nature of the threshold and how it might be non-arbitrarily generated, specification of the functions tu and au, whether or not real number measures are realistic, and even if they are, whether or not the values for au and for tu will be commensurate or not (as my heuristic talk of total utility suggests they might be); for it may be that moral value is a function of other things than utility for individuals and it may not be additive. Wessels does not stipulate otherwise, and indeed [Wessels, 2015] closes by noting the account leaves some questions open. But it does seem that her framework as is constitutes an important contribution to the literature on supererogation, and a structure that might accommodate it, and one of the few that employs formal modeling in the process of rich philosophical discussion, and so should be of further interest to both ethicists and deontic logicians.

Let me express my regret here that I have had to forego expositing and evaluating the body of work by Jan C. Joerden, along with Joachim Hrushka, at the interface of logic and supererogation.[101] This work shows a sustained effort to articulate logical expansions of the traditional deontic scheme, as well as more general discussions of various aspects of

[101] See [Hrushka and Joerden, 1987; Joerden, 1991; Joerden, 1998; Joerden, 2010; Joerden, 2012].

supererogation and allied concepts.

6.2 Åqvist's systematic frame constants models

The General Framework

Åqvist's approach to supererogation and kindred notions was inspired by his acquaintance with a version of [McNamara, 1993] prior to its publication[102], which led him to see unrealized deontic potential in some earlier more informal work on evidence [Åqvist, 1990]. Thus, McNamara's work can be used as a familiar bridge to Åqvist's framework. I will follow [Åqvist, 1999]), which is his main entry on the subject, although he mentions the problem of supererogation in [Åqvist, 2000].[103]

Åqvist relies centrally on a set of levels, but unlike McNamara who defines these as equivalence classes with respect to a quasi-ordering relation on the i-acceptable worlds themselves, resulting in equivalence classes of equi-ranked worlds constituting a partition of ordered cells. Åqvist instead begins with a partition of a set of worlds into a *finite* set of levels of acceptable worlds, with one more level added for all the unacceptable worlds (themselves not distinguished by levels). The finite levels are labeled 1 thru m (positive integers) and so derivatively, the levels, and their worlds are ordered. The levels/cells are designated $opt_1, opt_2, \ldots, opt_m$, constituting a partition of the set of worlds, W, in the structures that will be in focus. Note that the ordering is absolute, not relative to a world. These levels are understood as follows: opt_1 consists of the best (or optimal) worlds of W, opt_2 consists of the second best worlds– the best of $(W - opt_1)$, etc. However, opt_m is special in that it is the single cell containing all the unacceptable worlds, undistinguished from one another in values/levels. In contrast, all of the worlds in levels $opt_1, \ldots, opt_{m-1}$ are acceptable worlds, rankable derivatively according to their levels. Åqvist next stipulates a way to represent

[102]Letter from Åqvist to the author. "... I plan to do deontic logic your way...". Although a younger version of this author was very flattered and honored, Åqvist put his unique signature on the work.

[103]The underlying systematic frames framework is the subject of [Åqvist, 1997a], and it is briefly discussed in [Åqvist, 2002], but the topic of the framework's application to concepts outside of the usual ones covered in SDL (obligation, permission, etc.) and dyadic variants thereof does not arise in these places. [Åqvist, 1993a], covers a similar framework but there an ordering relation $\geq$ on worlds is used and the levels are more reminiscent of levels as characterized in much of McNamara's work (e.g. [McNamara, 1990], [McNamara, 1993]), but here Åqvist defines and labels the levels explicitly, $opt1, opt2, \ldots$ and tacitly assumes they are always at most denumerable in number.

such levels in the object language, with what he calls "systematic frame constants", $Q_1, Q_2, \ldots Q_i, \ldots$ (for $1 \leq i < \omega$), here just called "level constants" to stress their deontic/value-theoretic character. If one of these Qi denotes, it will be a level, some opt_i, and thus deemed to characterize that level uniquely (true at all worlds in that level and at no other worlds). He also adds universal modalities, N and M, for necessity and possibility respectively. That's it for the primitives of the language (other than the familiar truth-functional operators, $\rightarrow$, $\leftrightarrow$, &, $\vee$, $\neg$, $\top$, $\bot$, and parentheses).

More formally (cf. [Åqvist, 1999]), Åqvist first defines an *Hm-structure* as a four-tuple

> $M = \langle W, V, \{opt_i\}_{i=1,2,\ldots,m}\rangle$ such that
> 1) $W \neq \varnothing$ [W is a non-empty set of "possible worlds"]
> 2) V: Prop $\rightarrow$ $pow(W)$ [V is a valuation function which to each propositional variable assigns a subset of W]
> 3) $\{opt_i\}_{i=1,2,\ldots}$ is an infinite sequence of subsets of W
> 4) m is the positive integer "under consideration".

It is clear that what is intended is that for all $i > m$, $opt_i = \varnothing$. Åqvist now characterizes truth with the usual clauses for atomic and truth-functional operators, as well as these:

> $M, w \vDash \mathrm{N}\phi$ iff for each w' in W: $M, w' \vDash \phi$
> $M, w \vDash \mathrm{M}\phi$ iff for some w' in W: $M, w' \vDash \phi$
> $M, w \vDash Qi$ iff $w \in opt_i$ (for any positive integers i).

He then turns to what he calls "Hm models" as a "special kind of Hm-structures". I will call them the "H$'m$ Structures" (though his structures include valuation functions). Essentially, the H$'$m structures add constraints that assure the resulting frames fit the intended interpretation.

> H$'m$ *structures* are Hm structures satisfying these additional constraints (guaranteeing a partitioning of W):
> 1) For each i, j such that $1 \leq i \neq j \leq m$, $opt_i \cap opt_j = \varnothing$
> 2) $W = opt_1 \cup opt_2 \cup \cdots \cup opt_m$
> 3) For each i such that $1 \leq i \leq m$, $opt_i \neq \varnothing$.
> 4) For each i such that $i > m, opt_i = \varnothing$.

H$'m$ validity and satisfiability are defined in the usual way. The proof theory for any H$'m$ *logic* is the following:

(NEC) If $\vdash \phi$, then $\vdash \mathrm{N}\phi$
(AO) All tautologous formulas
(A1) S5-schemata: $\mathrm{M}\phi \leftrightarrow \neg\mathrm{N}\neg\phi$, $\mathrm{N}(\phi \rightarrow \psi) \rightarrow (\mathrm{N}\phi \rightarrow \mathrm{N}\psi)$, $\mathrm{N}\phi \rightarrow \phi$, $\mathrm{N}\phi \rightarrow \mathrm{NN}\phi$, $\mathrm{MN}\phi \rightarrow \phi$)
(A2) $Q_i \rightarrow \neg Q_j$, for all positive integers i, j with $1 \leq i \neq j < \omega$
(A3) $Q_1 \vee \ldots \vee Q_m$
(A4) $\mathrm{M}Q_1$ & $\ldots$ & $\mathrm{M}Q_m$.

A2 may be a little confusing at first glance, since there are an infinite number of positive integers, n, greater than any m in any H′m structure, but the idea is that all these opt_n must be empty levels (empty sets) and so all identical, and so taking any "two" of these, say n and r, $opt_n = opt_r = \varnothing$. But note that for any such $Q_n, Q_n \rightarrow \neg Q_r$ will be true (since $\perp$ implies anything), so A2 needn't be constrained by m; it can be stated in full generality. On the other hand, since each Q_i where $i \leq m$, uniquely specifies that i level (opt_i), then it will exclude any other such level denoter Q_j, $j \leq \omega$, empty or not. A3 and A4, restricted as they are to Q_i such that $1 \leq i \leq m$, match the intended interpretation. Since W is the union of all the m levels, A3 says at least one of the denoters of levels 1 through m holds; and for A4, since each associated level, $opt_i \leq m$, is non-empty, there will be a world in opt_i where Q_i holds, and hence $\mathrm{M}Q_i$ will hold.

Åqvist asserts that for any m, the Hm logic is sound, and strongly complete, and he cites a sketch of a proof contained in [Åqvist, 1997a].

Åqvist next introduces deontic operators of decreasing logical strength, as follows:

$$\begin{aligned}
&\mathrm{MU}_1 \stackrel{\text{def}}{=} \mathrm{N}(Q_1 \vee \cdots \vee Q_{m-2} \vee Q_{m-1}) \rightarrow \phi) \\
&\mathrm{MU}_2 \stackrel{\text{def}}{=} \mathrm{N}(Q_1 \vee \cdots \vee Q_{m-2}) \rightarrow \phi) \\
&\vdots \\
&\mathrm{MU}_{\text{m-2}} \stackrel{\text{def}}{=} \mathrm{N}(Q_1 \vee Q_2) \rightarrow \phi) \\
&\mathrm{MU}_{\text{m-1}} \stackrel{\text{def}}{=} \mathrm{N}(Q_1 \rightarrow \phi).^{104}
\end{aligned}$$

Note that Q_m is missing, else MU_1 would just be essentially N, but since Q_m is interpreted as containing *all* unacceptable worlds, MU_1 is essentially McNamara's **M**U— true for whatever holds in all acceptable worlds, and $\mathrm{MU}_{\text{m-1}}$ is essentially McNamara's **O**U— true for whatever holds in all the best acceptable worlds. However, since Åqvist has a denumerable number of sequential levels and likewise for level constants,

[104] In [Åqvist, 2000], he labels the definientia O_1A, O_2A, $\ldots$ $\mathrm{O}_{m-1}A$.

all the levels between the best acceptable level and least acceptable level are denoted in the object language, and they must be finite in number. MU_2 is true of whatever is entailed by all the acceptable worlds minus those ranked least of the acceptables, and MU_3 is true of whatever is entailed by all but the least and second to least acceptables, and so on until we have what is entailed by the best levels. A simple mnemonic here is to think of MU as a necessity operator *function* taking a positive integer as argument, and yielding a necessity operator keyed by its subscript to a specific set of worlds. Subscript 1 "shaves off" the very lowest level of worlds, namely opt_m the unacceptable worlds, then subscript 2 shaves off the lowest two levels, the unacceptable level and the level with the least acceptable worlds, and then we keep shaving off one more as we ascend the numbers until we have all levels shaved away except for the best, $MU_{m\text{-}1}$.

Åqvist then introduces a corresponding operator to be interpreted as a multiply ambiguous notion of "Wrong", defined in the usual way, as essentially, "Mustn't":

$$WR_i\phi \stackrel{\text{def}}{=} MU_i\neg\phi. \text{ [WRONG}_i\phi \text{ iff MUST}_i\neg\phi\text{]}$$

Application to supererogation and kindred notions

Åqvist initially tests his system against three supererogation cases drawn from those of McNamara[105] In the first case, he imagines we have only 3 levels, and so he invokes an H3 logic where the best level involves an act of supererogation, the second best (the least good of the acceptables) involves another acceptable performance, and the lowest level is the collection of unacceptable worlds. Where R is "Passerby Jane *rescues* an infant from a dangerous fire", he then stipulates the following interpretation of claims about R's status:

> MU_1R as It *must* be that R [R holds in all acceptable levels]
> MU_2R as It is *good* that R [R holds in all acceptables levels other than the least acceptable one]

As the case is conceived, $\neg MU_1R$ & MU_2R holds, for there is a lowest acceptable level without her rescue (thus $\neg MU_1R$) but shaving off the

[105]He cites [McNamara, 1993], and two match of the three he uses, but the same or very similar cases occur elsewhere, for example, [McNamara, 1990; McNamara, 1996b; McNamara, 1996c].

least acceptable level (level 2) takes us to the best level, where the rescue occurs (thus $\mathrm{MU}_2\mathrm{R}$). Given the proposed interpretation, Åqvist suggests this fits Chisholm's often used gloss for supererogation, "Non-obligatory well-doing:

(1) $\neg MUST(\mathrm{R})$ [for $\neg\mathrm{MU}_1\mathrm{R}$] and (2) $GOOD(\mathrm{R})$[for $\mathrm{MU}_2\mathrm{R}$]

The analysis of *good* proposed here seems problematic. It can surely sometimes be good to do what is required even by doing the least one can do.[106] For example, suppose our agent is now not a passerby but a fire fighter on duty, yet there nonetheless might be two options for how to perform the rescue, one the minimum required of the firefighter, one beyond the minimum because though faster, it is much riskier. Even given the role of firefighter, both can be highly dangerous and beneficial. Would it not be good to do even just the minimum slower less dangerous rescue even if another alternative rescue is even better though also not required? But this rescue will not occur among the worlds above the minimal level, so it can't be good on the proposed analysis.

Åqvist then turns to a variant H3-logic case where we imagine there are two acceptable levels again, the best one where the agent *rescues both* of two infants, Jill and Bill, in danger (in different parts of the building); the next best level where she rescues *just* Jill or *just* Bill, and here he imagines rescuing at least one is required, so the unacceptable level involves no rescue. With obvious abbreviations for rescuing Jill and rescuing Bill, he suggests the following mapping to our pre-theoretic notions

(1′) $\neg MUST(\mathrm{B}\ \&\ \mathrm{J})$ [for $\neg\mathrm{MU}_1(\mathrm{B}\ \&\ \mathrm{J})$]
(2′) $GOOD$ (B & J) [for MU_2 (B & J].
(3) $\mathrm{MUST}(B \vee J)$ [for $\mathrm{MU}_1(B \vee J)$

He takes the first two clauses to apply, rescuing both is non-obligatory but good and so supererogatory, and the last clause means it is unacceptable to not rescue at least one. A similar problem seems to arise here for *good*.

His final case is an H4 case where now three infants are trapped (Bill, Jill, and Phil), and it is obligatory to rescue at least one, but possible and permissible to rescue exactly two or all three. In the best level, all

[106] See the solider on point example in [McNamara, 2011a] and [McNamara, 2011b].

three are rescued, in the second best, some two only of the three are rescued, and in the third best you rescue one only, and in the unacceptable level, you rescue none. He then suggests this interpretation of three pre-theoretic notions in a 4-levels case, the pre-theoretical interpretations as indicated in the left column, with the proposed analysans in brackets on the right:

(1)	$MUST$(J $\vee$ B $\vee$ P)	[for MU_1(J $\vee$ B $\vee$ P)]
(2)	$\neg MUST$((J & B) $\vee$ (J & P) $\vee$ (B & P))	[for $\neg MU_1$((J & B) $\vee$ (J & P) $\vee$ (B & P))]
(3)	$OUGHT$((J & B) $\vee$ (J & P) $\vee$ (B & P))	[for MU_2((J & B) $\vee$ (J & P) $\vee$ (B & P))]
(4)	$\neg OUGHT$(J & B & P)	[for $\neg MU_2$(J & B & P)]
(5)	GOOD(J & B & P)	[for MU_3(J & B & P)]

At the semantic level, the first says that at all acceptable levels someone is rescued, the second says that that is not so for rescuing any pair, the third says that at all levels above the minimally acceptable level (opt_3), at least two are rescued, the fourth says that that is not so for rescuing all three, the fifth says that at the acceptable levels other than the minimally acceptable one and the next best one (which in an H4 structure is the best level), all three are rescued. Although (1) and (2) are fine, the proposed analyses and applications of *ought* and *good* here are problematic. First it is even less clear as to why *good* in an H_4 context applies only to options that occur in the best level. Surely it can be good to rescue even just two in natural scenarios fitting the case, and it might be argued that it must be good to do more than the least you are permitted to do (as in this case) Also, notice that we get the odd result that it is not the case that you ought to rescue all three, but this is meant to be a deontic/moral ought, and it seems more plausible to say you *ought* to (though you don't *have* to) do the *best* you can here, the contradictory of the left side of (4). McNamara would identify *ought* with whatever permeates the best level (opt_1), so the condition proposed for *good* in (5) seems better suited to *ought*.

Åqvist then quickly suggests that we can easily transpose the prior cases and the associated analyses to capture permissive ill-doing using negation and a set of proposed negatively valenced identifications of pre-theoretic normative statuses with those of his deontic system. This, he suggests, would find a place for what is *bad* but permissible, or called an "offence" by Chisholm and Sosa (often now called "suberogation"). He says that for each of the three cases analyzed above

> "In the statements characterizing the case at issue, replace each operator MUST_i $(i = 1, \ldots, m-1)$ by the matching compound operator WRONG_i" [Åqvist, 1999, p. 270]

He here obviously intends that we add a "¬" after the latter operator in the replacements. For example, regarding the first case, Åqvist suggests that ¬R (Jane's not rescuing) is permissible, but "bad (in the sense of having a good negation)". [Åqvist, 1999, p. 271], so he suggest that we have "permissible ill-doing" represented, Chisholm's (and Sosa's) oft used gloss for an "offence". For brevity here, I consider the conversions only for the first case above. The conversions are on the left side, with the equivalences that constitutes their backings in brackets:

Case 1 conversion equivalences:
1) $\neg MUST(\mathrm{R}) \leftrightarrow (1')\ \neg WRONG\neg(\mathrm{R})$
[since $\neg\mathrm{MU}_1\mathrm{R} \leftrightarrow \neg\mathrm{WR}_1\neg\mathrm{R}$)]
2) $GOOD(\mathrm{R}) \leftrightarrow (2')\ BAD\neg(\mathrm{R})$ [since $\mathrm{MU}_2\mathrm{R} \leftrightarrow \mathrm{WR}_2\neg\mathrm{R}$]

Åqvist provides the right side of the equivalences on the left column (with the primes after the numerals). I place in brackets the associated equivalences in his formal system. He then proposes that for Case 1, that in an H_3 system (so semantically, in a context where there are only 3 levels of worlds), the conditions proposed for analyzing $(1')$ ¬WRONG¬(R) and $(2')$ $BAD\neg(\mathrm{R})$ together illustrate the notion of what is permissible but bad, or permissible ill-doing, and thus Chisholm & Sosa's gloss for an *offence* (here an offence of omission).[107] So it is an offence that the agent does not rescue the child from the dangerous fire. We have similar proposed conversions for Case 2 and Case 3: for Case 2, it is permissible but bad and so an offence of omission for the agent to not rescue both Bill and Jill from the dangerous fire; and for Case 3 (with 4 levels), it is permissible but bad to fail to rescue all three potential victims, and so an offence of omission, even though here rescuing even two appears to be beyond the minimum required (to rescue one).

Let me note that, first of all, we have a similar problem here in the proposed analysis of *bad* that we saw for *good*. Based on the structure and story of Case 1, there is no way to tell if it is positively *bad* to not perform the dangerous rescue, perhaps the minimum option itself can involve risk or sacrifice. The proposal amounts to saying it is *bad* to

[107] So the condition proposed for $\neg WRONG\neg$ is also offered as analysis for what is just plain *permissible*.

not do the very best, no matter how heroic the best is.[108] In the third case, of the type that McNamara uses to illustrate that one can have two options each beyond the call (so permissible), but one better than the other, doing anything other than the best is deemed permissible but bad, so that it is deemed an offence for the passerby to do the dangerous rescue of just two of the children.

At this juncture, Åqvist begins to reflect on the three cases, and move toward generalizing, but it needs to be understood that what we have here is an unending series of deontic logics, so generalizations will be across systems not within one system. He also offers analyses of a series of notions of indifference, one essentially optionality, the strongest essentially McNamara's analysis of *indifference*, along with a series of notions of strength in between these depending on the number of levels beyond 3 in the semantics. He likewise generalizes his analysis of obligation, permission, prohibition, supererogation, offence, leading to an enriched classificational scheme (beyond TTC), depending on the number of levels associated with the number-indexed logic, thus proposing a solution to "the problem of supererogation" as described by [Chisholm and Sosa, 1966a]. He is then able to prove increasingly larger classificational schemes (mutually exclusive and jointly exhaustive normative positions) depending on the number of levels (or equivalently, the number-index of the logic).

In the interests of space, I pass over most of these here for another occasion, and only make explicit the analyses of supererogation and offence, *relative to* H_3 logics:

$$\mathrm{SUP}(\phi) \stackrel{\text{def}}{=} \neg\mathrm{MUST}(\phi) \;\&\; \mathrm{GOOD}(\phi) \text{ [i.e. } \neg\mathrm{MU}_1\phi \;\&\; \mathrm{MU}_2\phi]$$
$$\mathrm{OFF}(\phi) \stackrel{\text{def}}{=} \neg\mathrm{WRONG}(\phi) \;\&\; \mathrm{BAD}(\phi) \text{ [i.e. } \neg\mathrm{MU}_1\neg\phi \;\&\; \mathrm{MU}_2\neg\phi]$$

Reflections on supererogation and kin in this framework

As is typical of Åqvist work, this is something of a technical tour de force. He is able, by shifting to explicitly named levels, to construct necessity and possibility operators for each level and generalize standard techniques to get determination theorems. One cannot help but be impressed. The framework allows, at least to a first approxima-

[108] Similarly results follow for Case 2, where not rescuing both is deemed *bad*, and still worse for Case 3, where it is bad to not rescue all three, so it is bad to rescue say Bill and Jill but not Phil, even though that involves doing more than the minimum acceptable.

tion, a large enough set of distinctions that one can at least consider proposing to distinctly represent some of the neglected moral statuses of the traditional scheme, such as supererogation and indifference and kindred notions. It also allows Åqvist to apply the framework to analyzing conditional obligations [Åqvist, 1993a; Åqvist, 1997a; Åqvist, 2000; Åqvist, 2002], impressive work we must pass over. This is no small achievement.

That said, the framework appears to face some serious challenges. We noted a few above in passing while expositing his analyses of the three cases, but let me step back and note some more general problems. One is that we get a hierarchy of logics, rather than a single unified framework. Is that desirable over a unified logical framework? Second, the semantics presupposes that there is always a finite number of permissible levels of value in any frame, but this seems an unattractively narrow assumption. Perhaps I have a dial that allows me to produce say pleasure (or reduce pain) for a given individual. Why should we assume that, all else equal, this will not provide for a continuous subset of levels, infinite in number, perhaps denumerable (it turns in clicks, but from 1, $\frac{1}{2}$, $\frac{1}{4}$, etc.) perhaps non-denumerable and truly continuous. Suppose we were to modify the semantic framework (as in McNamara's DWE semantics) so that there is possibly a denumerably infinite number of distinct non-empty levels of value, and we simply designate the lowest ranked level as 1 say, making other needed adjustments. This underscores a third problem. If we are viewing this system as offering a representation of our own moral concepts, don't we now have an unlearnable (object) language with an infinite number of *primitive* terms? How can we learn the meaning of $Q_1, \ldots Q_n, \ldots$ where infinite?[109] It is said that it is modelled on the Andersonian-Kangerian reduction, but there we have a contentful reading for the constant/s used, and so no problem conceptualizing them (see [McNamara, 1999], [Åqvist, 2002]), for example d as *all of morality's demands are met*; but how do we conceptualize $Q_1, \ldots, Q_m, \ldots$? Even if a large finite sequence, it is not clear how we can get onto exactly what characterizes each level. The meaning would seem to change with the situation. If we read say Q_i as "the ith level of value obtains" we must assume there is no continuous case, else we need

[109]When I first presented the DWE framework ("Doing Well Enough: Toward a Logic for Common Sense Morality." Society for Exact Philosophy conference, York University, Toronto, Canada, 1993), Daniel Bonevac asked in Q & A if I had considered having a constant for each level, much like Åqvist, but since I thought it was ad hoc to think they would be finite in number, I replied that I did not want to use a formal approach that would appear to make the language unlearnable.

a non-denumerable set of constants. Fourth, Åqvist speaks as if moral terms are multiply ambiguous, with *unlimitedly* stronger and weaker senses of *must, wrong, indifference*, etc.; but is this at all plausible? It cuts against the main grain of current research on modal auxiliaries and kindred terms, and against the common sense perception that even if there are *some* ambiguities, they are not limitless, ever increasing as the number of levels of value increases. Fifth, Åqvist seems to see himself as articulating the sort of framework that Chisholm and Sosa desired, but he seems to not notice that he endorses some of Meinong's theses that Chisholm and Chisholm-Sosa are at pains to rightly reject,: SUp $\leftrightarrow$ OF$\neg p$, and OFp $\leftrightarrow$ SU$\neg p$, as we saw in our exposition of Chisholm's landmark [Chisholm, 1963b] in Section 3. (Are all supererogatory heroic rescues *bad* to not perform?) As we also saw in Section 5, these equivalences run contrary to the classical conception of supererogation as something optional and good or praiseworthy to do and *not* bad or blameworthy to not do.[110] Such equivalences hold for quasi-supererogation and quasi-offence, so at best we would have these notions analyzed, not the target ones. Åqvist seems to conflate what McNamara calls *permissible suboptimality* with *blameworthy permissible suboptimality* (see Section 5). Sixth, I think Åqvist's analysis of *good* and *bad* are flawed as well. Let's focus on *good*. He ultimately analyzes what is good as always excluding the minimally acceptable options. But McNamara's soldier on point example counts again this analysis [McNamara, 2011a; McNamara, 2011b]. Sometimes, obligations can be just so arduous that even doing the least you can do (fulfilling the obligation in the minimally acceptable way) is a good thing to do, even if there are still better options. On Åqvist's account, doing the least you can do in discharging an obligation is never good, no matter how arduous the obligation, no matter how many of us would shirk that duty if faced with it. Lastly, as we saw, Åqvist's analysis of *ought* has implausible consequences. It entails that if there is a less than best level that is not a minimal level, than it is false to say you ought to do whatever is best for you to do. Sometimes we ought to do more than just anything better than the least we can do. "You ought to do the best you can" seems very plausible even if the analysis of what is best is not what the act utilitarian says it is. (Contrast that with "you *must* do the best you can".)

Perhaps better applications of Åqvist framework might yield better results — some of the problems may be misapplications of the frame-

[110] As we saw with Chisholm's initial framework, Åqvist thesis that $GOOD(A) \leftrightarrow BAD(\neg A)$, and variants is not in keeping with Chisholm as we saw in Section 3 (nor with Chisholm-Sosa's accounts mentioned earlier).

work, but there seem to be more fundamental problems. We have a disunified series of logics rather than one logical framework. And we seem to have something of a dilemma: he just assumes there must always be a finite number of acceptable value levels (each represented by a constant in the object language); and if we switch to allowing an infinite number to fix this, then the analytic machinery seems to yield an unlearnable language with an at least denumerable number of primitive constants naming the levels.

6.3 Sven Ove Hanson's work and supererogation

Deontic logic, orderings, and making room for supererogation

Sven Ove Hanson has produced a high quality body of work in deontic logic. A central portion is focused on addressing challenges that SDL faces (and its classical ideal worlds semantic conception). He does this by specifying various preference relations on propositions (or formulae), and then defining *various categories* (or *types*) of operators (or predicates) via relationships between operator applications and ordering relation constraints. He then attributes various de facto deontic operators to such operator categories.[111] A closely related project (or better, a facet of a larger project) is to similarly make use of preference orderings to analyze an array of operator categories for evaluative notions.[112] In this sense, the two are unified, as indicated in his book, [Hansson, 2001], where a very impressive array of categories are defined (newly defined to the best of this author's knowledge) and systematized via (in turn defined types of) preference relations. This is done at a high level of generalization and systematization.

Let me illustrate Hansson's approach briefly by exploring various deontic operator categories that Hansson defines vis a vis the interactions of their instances with one particular ordering relation.[113] Assume we have a preference relation, $\geq$, in our object language, one that is transitive as well as being *interpolative*–a constraint on the preference relation vis a vis disjunctions to the effect that either the disjunction is

[111] See especially [Hansson, 1990b; Hansson, 2001; Hansson, 2004], and in the first volume of this handbook, [Hansson, 2012].

[112] See especially [Hansson, 1990a; Hansson, 2001].

[113] Here I will follow [Hansson, 2012] as much as possible, but will indicate when I venture elsewhere. Note that Hansson often casts things via *predicates*, and an ordering relation on *sentences*, but I recast via *operators*, and assume the ordering is over *propositions* (e.g. as sets of worlds). This should still convey the basic picture well enough.

equi-ranked with one of its disjuncts or it is intermediate in value between them.[114] Now add that we have various operators $O_1, \ldots, O_n$, and then we can define some operator categories (or classes) via conditions they must meet regarding the paired preference relation, such as:

an operator O_i is $\geq$-*positive* iff $(\mathrm{O}_i\phi \;\&\; \psi \geq \phi) \rightarrow \mathrm{O}_i\psi$
an operator O_i is $\geq$-*negative* iff $(\mathrm{O}_i\phi \;\&\; \phi \geq \psi) \rightarrow \mathrm{O}_i\psi$.

So an operator is *positive* (that is $\geq$-positive–assume the "$\geq$-" qualifier is intended throughout) iff it applies to any proposition ranked at least as high as one it applies to, and it is *negative* iff it applies to any proposition equi-ranked or out ranked by one it applies to. Hansson suggests that some evaluative operators such as *it is good that, it is best that, it is not bad that*, ought to be positive if faithfully represented in a formal language. Similarly, for *negative* operators intended to faithfully represent notions like *it is bad that, it is worst that, it is not good that.* How might deontic operators fit in? Here we will concentrate of the traditional core cases intended for modeling in SDL. He suggests that **P**E, if intended to represent permissibility, should be represented as positive, for if it is permissible that ϕ and ψ is ranked at least as high as ϕ, then it is permissible that ψ. What of **O**B? In anticipation of his answer, let us add two more of his operator categories:

an operator O_i is $\geq$-*conrtrapositive* iff $(\mathrm{O}_i\phi \;\&\; \neg\psi \geq \neg\phi) \rightarrow \mathrm{O}_i\psi$
an operator O_i is $\geq$-*contranegative* iff $(\mathrm{O}_i\phi \;\&\; \neg\phi \geq \neg\psi) \rightarrow \mathrm{O}_i\psi$.

So an operator is *contrapositive* iff it applies to any proposition (ψ) whose negation ranks at least as high as the negation of one (ϕ) that it applies to, and it is *contranegative* iff it applies to any proposition (ψ) whose negation is ranked at least as low as the negation of one (ϕ) that it applies to. Then assuming the familiar SDLish interdefinablility equivalences between **O**B, **P**E, and **I**M (e.g. $\mathbf{P}\mathrm{E}\phi \leftrightarrow \neg\mathbf{O}\mathrm{B}\neg\phi$), he notes that there are tight relationships between answers to the question of how to categorize these three deontic operators by showing the following equivalences (given the ordering properties cited):

[114] That is, either $(\phi \vee \psi) \approx \phi$ or $(\phi \vee \psi) \approx \psi$ or $\phi > (\psi \vee \phi) > \psi$ or $\psi > (\psi \vee \phi) > \phi$, with $>$ and $\approx$ defined in the usual way via $\geq$. It follows from the two properties that the relation is also complete (connected).

(a) **O**B is $\geq$-contranegative iff *(b)* **P**E is $\geq$-positive iff *(c)* **I**M is $\geq$-negative
(d) **O**B is $\geq$-positive iff *(e)* **P**E is $\geq$- contranegative iff *(f)* **I**M is $\geq$- contrapositive

Consider the first equivalence triple *a-c*. He notes that although the positivity of **P**E is linked to the negativity of **I**M in ways the reader might expect, it does not treat **O**B as positive (even though as indicated above, *many different* operators are intended to, and appear to, have properties like positivity). Why? For one, a positive categorization would not make room for supererogation. An option might be *strictly better* than an obligatory one and yet be supererogatory, and hence not obligatory. For example, it might be better for Jane to be such that she rescues the child from the building than to be such that either she rescues the child or (instead) pulls the fire alarm down the road and waits to direct the fire trucks.[115] So a faithful representation of obligation will not be as an operator that is $\geq$-positive (given a natural reading of $\geq$), else supererogation will be ruled out.

Above we compared cases of an obligation with different ways of fulfilling it. One potential problem about classifying obligation as contranegative is that in the case of something supererogatory, there will be cases where we have a minor obligation, say to return a book to a friend *(b)*, and quite unrelated to this, we have say a momentous supererogatory rescue of a child from a fire *(r)*. Here we imagine that although b is obligatory and r is not, nonetheless, r is ranked much higher than b, and, $\neg b$ is ranked much higher than $\neg r$ (failing to fulfill the small obligation is better than not making the supererogatory rescue) — note there is no forced choice here, just a point of comparison. Since there can clearly be minor obligations and major acts of supererogation, taking obligations to be contranegative requires saying the violation of such minor obligations is always worse than not living up to a supererogatory

[115]To cast things in terms of McNamara's rescue example discussed above, the fire alarm pulling, which precludes Jane's making a rescue, is the minimal way to discharge the duty, and so the background assumption operating here is perhaps that a proposition describing an obligation is ranked at the levels of the least one can do in discharging it. For the disjunction above, that in turn suggests that the value of the disjunction is ranked equally with its lowest ranked disjunct. It seems that if we are too maintain McNamara's insight about going beyond the call as doing more than the least one can permissibly do, then there is pressure to also rank disjunctions as equi-ranked with the minimal ranked disjunct.

ideal.[116] This is a tentative probe.[117]

This is but a brief representative sample of the logical richness of Hansson's framework, one naturally skewed toward our chapter topic. Hansson uses this framework to provide alternatives to SDL that avoid some of the paradoxical results, especially those associated with the rule RM, as well as exploring various conceptions of evaluative notions like those above for good, bad, and indifferent, and a number of other things beyond this chapter's scope. The consolidation of much of his work on preference and deontic logic in [Hansson, 2001] is, in this author's opinion, a tour de force and deserves more attention that it has received.[118]

On "Representing Supererogation"

In a recent short survey article [Hansson, 2013], Hansson discusses what he sees as limits of prior approaches to the logic of supererogation, and he provides the outlines of a proposal of his own. We here assess this survey (briefly) and primarily focus on his proposal. However, before doing so, let's note that although Hansson takes Chisholm's gloss on supererogation (non-obligatory well doing) to be insightful, the upshot of the criticism that McNamara provides of the traditional analysis ($\mathbf{S}\mathrm{U}\phi$ & $\mathbf{P}\mathrm{W}\phi$ & $\neg\mathbf{B}\mathrm{W}\neg\phi$) applies here as well: in demanding situations, it can be good to do even the least one can do (where one could permissibly do even better), and Hansson himself points out a general problem of *trivial variants* to show these sorts of analyses are not adequate:

> Suppose that I am morally required to visit Ms. X and apologize for some problems I have caused her. Doing so is a good (and praiseworthy) action. Now consider a trivial variant of that action, namely to visit her and apologize, entering her apartment with my left foot first (p). This is clearly a good (and praiseworthy) action, but it is not obligatory since I might just as well have entered with my right foot first. (p. 8)[119]

[116] Or saying they are simply incomparable in all such cases, which would need justification, and would not fit the current ordering properties.

[117] It might be argued that $\neg b$ is blameworthy, but, as in the classical conception, r is not, and any blameworthy option should be ranked below any non-blameworthy option; furthermore, some think blameworthiness entails impermissibility. I am grateful to Sven Ove Hansson for suggesting a reply along these lines.

[118] Compare [Arlo Costa, 2003] and [Horty, 2002], who provide brief critical overviews coupled with high commendation.

[119] All lone page references in this section will be to [Hansson, 2013].

The latter point is used to motivate his own positive account, since the solution proposed involves the notion of a variant, in particular that of *p being a better variant of an obligatory q* (p. 8). Hansson thus endorses supererogation *as fundamentally relational* (Cf. Wessels in 6.3), specifically as *oversubscription* to a particular obligation/duty.[120]. "p is a variant of q" is here provisionally analyzed as syntactic entailment, $p \vdash q$ (or perhaps proper syntactic entailment by adding $q \nvdash p$). It is also clear that Hansson is thinking of p and q here as two "action representations" (e.g. p. 9). So supererogation here is analyzed as a dyadic notion, say $\mathbf{S}_{\mathrm{p}}q$, read as p is supererogatory with respect to q, with this as analysans:

$$p \vdash q, Oq, \neg Op, \neg O\neg p, \text{ and } p > (q \;\&\; \neg p)$$

Hansson's informal reading of the proposed analysis is that "a supererogatory action is an optional action that is a better variant of another, obligatory action" (p. 9). Obviously, the set of conditions above cannot be formed into a conjunction as it stands, given the nature of the first element, so let's introduce a necessity operator, and replace the set above with a genuine conjunction, state the analysis as follows, and shift to some of the notation we have been using previously:

$$\mathbf{S}_p q \stackrel{\text{def}}{=} \Box(p \rightarrow q) \;\&\; \mathbf{O}\mathrm{B}q \;\&\; \mathbf{O}\mathrm{P}p \;\&\; (p > (q \;\&\; \neg p)).$$

Hansson also indicates that the obligation operator is to be read as an all things considered operator, but for reasons already discussed (e.g. to rule out unresolvable conflicts), let's read it more strongly as expressing what is overridingly obligatory.[121]

Hansson goes on to offer a negative analog to supererogation, which he calls a *substandard act*, which we will symbolize as $\mathbf{S}'_p q$. He offers this set of conditions as analysans:

$$p \vdash q, \mathbf{O}\mathrm{B}q, \mathbf{O}\mathrm{P}p, \text{ and } (q \;\&\; \neg p) > p,$$

and for the same reasons just noted, we express this proposed analysis this way:

[120] Cf. [Feinberg, 1961], though it is of course linked to the etymology of the term– roughly, to *over pay out*.

[121] Hansson defends $(\mathbf{P}\mathrm{E}q \;\&\; p \geq q) \rightarrow \mathbf{P}\mathrm{E}q$, and so if q is overridingly obligatory, then it is permissible, and so we could reduce the third conjunct to just $\mathrm{PE}\neg p$, dropping $\mathbf{P}\mathrm{E}p$ since then entailed by the 2nd and 4th conjuncts.

$$\mathbf{S}'_p q \stackrel{\text{def}}{=} \Box(p \rightarrow q) \;\&\; \mathbf{O}\mathrm{B}q \;\&\; \mathbf{O}\mathrm{P}p \;\&\; ((q \;\&\; \neg p) > p).$$

He says of this notion: "Substandard variants of obligatory actions do not seem to have been discussed previously, but they are common enough in everyday moral discussions." (p.10) He then says in defense of the uniqueness claim:

> "They should be distinguished from suberogatory actions. This is a category that was introduced by Chisholm [3] under the names of 'offence' (p. 2) and 'permissive ill-doing' (p. 5) and renamed 'suberogatory acts' by Zimmerman [30, p. 375]. McNamara [22, p. 155] uses the term 'permissibly suboptimal' to denote actions that are permissible but not implied by what morality recommends, i.e. $Pp \;\&\; \neg Op$ which is equivalent to $\neg O\neg p \;\&\; \neg Op$ if permission and obligation are interdefinable in the standard way. Therefore, if p is substandard with respect to some q, then p is permissibly suboptimal."

This is confusing. The negation in front of the first occurrence of "O" above is misplaced. As we have seen, McNamara uses "permissible suboptimality" for the condition $\mathbf{P}\mathrm{E}p \;\&\; \mathbf{O}\mathrm{U}\neg p$, so that what morality recommends *rules out* p, that is, $\neg p$ is implied. Also, given McNamara's semantics, p occurs only below the optimal range of complete alternatives, so if the agent acts permissibly and p occurs, then there is an obligation such that permissibly fulfilling it conjointly with $\neg p$ occurs only at complete alternatives outranked by the best p alternatives. So the account has stronger affinities to McNamara's account of suboptimality, which is *not* identified with that of an offence by McNamara for the reasons we saw earlier linked to the need to weave in agent-evaluative concepts. Indeed, the difference McNamara asserts between suboptimality and suberogation has affinities to just what Hansson thinks differentiates his view from that of someone like Chisholm's. There is a genuine difference between Hansson's analysis of suberogation due to its explicit relational character. So, like the account offered of supererogation, what is essentially different here is that the basic account is dyadic, some action is suberogatory or substandard *only relative* to some particular obligation, and then any monadic notion of suberogation would be analytically derivative presumably.

Turning back to supererogation, note that the analysis seems to only make sense if we are assuming that an obligatory action should always

be appraised as having the same value as the value of its minimally permissible way of fulfilling it (to use McNamara's terminology), since it would seem that you fulfill the obligation in a supererogatory way as long as you fulfill it in a way that produces more value than the minimal way of doing so involves. If the bar for the value of the relevant obligation is *higher than* the minimum involved in fulfilling the obligation, then we will be left with cases where one goes beyond the minimum that are not treated as beyond the call. We will assume going beyond the minimum suffices, and set aside for the moment how we will integrate this with how we are to conceptualize the value of all variables (e.g. even impermissible ones). Note also that when we say p is better than q-without-p, we are not speaking only from the standpoint of this particular obligation (e.g. not "considering only the obligation q, p is a superior way to fulfill that one"), for the ranking is of propositions *generally*, as the notation reflects: $p > q$.

These interesting analyses raise a number of questions. Hansson himself considers two objections, framed with regard to the analysis of supererogation. The first is that it is objectionable that the very same p could be supererogatory relative to obligation q, but not to obligation r, and so there is no sense given for when something is supererogatory simpliciter, but of course many actions (if not all) that are supererogatory in some sense are supererogatory per se. Hansson provides a case to illustrate (pgs. 6-7) where briefly, he promises to *give a special book to an elder relative (r)*, and of those elders, he has two, an aunt and an uncle. It so happens that it is his aunt's birthday, and *he is obligated to provide her with some gift (q)*, so *he gives her the book (p)*, but it is also the case that his uncle, who is impoverished, would enjoy the book much more than his aunt. Hansson says of p: "as a variant of q it was clearly supererogatory; it was optional for me to do it ... and it was clearly better than not doing so $(p > (q \,\&\, \neg p))$"(p.11), but he also adds that $(r \,\&\, \neg p) > p$ (since this entails giving the book to the uncle), so $\neg \mathrm{S}_p r$ per the analysis. This is meant to bring out the objection: there is no account of supererogation simpliciter here, with which Hansson agrees, but he suggest tentatively that perhaps a derivative account of a monadic conception is possible.[122] We will return to this, but first let's

[122]Hansson tentatively proposes a possible analysis of monadic supererogation via his dyadic notions (taken as basic), in the spirit of his proposal, the monadic notion might be analyzed this way, he suggests: $\mathbf{S}_p \stackrel{\text{def}}{=} \exists q \mathrm{S}_p q \,\&\neg\exists q \mathbf{S}'_p q$. That is, an act is supererogatory per se iff it is a supererogatory variant of at least one obligation and a suberogatory variant of no obligation. We leave discussion of this interesting proposal for another occasion.

probe this case a bit. Working backwards, the latter is plausible, since in the circumstances, it is unalterable for him that he can fulfill r while not giving his aunt the book (p) only by giving it to his uncle and since his uncle is destitute, and will get more out of the book, giving it to the uncle seems better than giving it to the aunt. Given the intentionally relativized sense advocated for supererogation here, the case supports saying that fulfilling q by p was supererogatory, but not it seems for the reasons stated nor per the analysis. For is it really "clearly better than not doing so ($p > (q \ \& \ \neg p)$)" He can fulfill his obligation to give his aunt a present (q) in many ways other than p, and it sounds like he can do something really special by doing r, and so not p; in which case, why should we say that giving that particular book to his aunt is better than giving her a gift but not that book? I think this case reveals that the analysis is not doing what is intended. For that, we would in some sense need multiply relativized ordering operators of the form $>_q$, and $>_r$, etc. essentially one for each obligation. For given the case description, giving the book to the aunt can only be better than giving her a gift without that being the special book if we ignore the relevance of r and the value of the book to the uncle. Given the whole picture, it is not at all obvious why the converse does not hold, $(q \ \& \ \neg p) > p)$. I suspect what we have here is not so much a relativized notion of supererogation as the monadic notion with a tacit operator prefacing it like *considering only the obligation q (and perhaps some limited background info)*, p is supererogatory.[123] So there seems an instability in the account given and how the particular-obligation-relativized conception of supererogation will be reconciled with a non-obligation-relativised ordering relation.[124]

Hansson considers a more direct challenge to his account, namely the contention that there can be a supererogatory action without the action constituting an oversubscription of any obligation (i.e. that the act is a variant of). He gives a nice example of seeing a stranger in distress and stopping to talk for some time and calm the person. In reply, Hansson essentially takes the position of ethicists who want to say that supererogatory acts are always fulfillments of general obligations, often so-called imperfect obligations, like an obligation to be kind or decent.

[123]Compare *considering only that I promised to have lunch with my friend, I am obligated to do so.* Here there is no reason to think we have a dyadic notion of obligation operating.

[124]There is a conception of supererogation than can be viewed as relativized without this problem: doing more good than one was obligated to do, but here presumably the amount of good to be exceeded will be determined by the minimum required by the totality of one's overriding obligations combined, as in McNamara's account.

This looks like a very substantive position to take for a logic, since rather controversial even in ethical theory, and there are challenging problems especially with the conception of the fulfillment conditions for an imperfect obligation, like being kind. The variant must also count as fulfilling the associated obligation if it is to be a variant of it. But is the obligation to be kind *fulfilled* by stopping to talk to the stranger this one time? Surely it is kind to do so (let's assume compassionate motives), but is that enough? If the obligation is be kind *now*, then we have moved even more strongly into the substantive realm and risk endorsing a rather demanding ethics from our logic (e.g. *be kind whenever a good opportunity arises* is alas, relentless in our world full of destitution).

Hansson notes as "particularly problematic" cases where there appears to be no obligation that a paradigmatically supererogatory act is a variant of, for example, where someone risks her own life to save someone else, but there is nothing short of this that she could have done to help the person in danger" (p.12). He goes on to remark that this is a quite unusual case and "supererogation," if applied, derives from the more ordinary cases he discusses. It is not clear what the force of this reply is. Whatever its origin (e.g. overpaying the innkeeper in the good Samaritan parable), if the term applies, and there is no ambiguity, then it would seem the conclusion is that the proposed analysis does not apply to all cases of supererogation, because they are not all cases of oversubscription to an obligation — fulfillment of a particular obligation via a variant thereof. This is my own view, that oversubscription is a general but not exhaustive case, and that the tie to obligation consists in doing more than you would have done if you had fulfilled your overriding obligations in the minimal way; but here, note that it is *the plurality* of one's strict obligations, not a single one, that is central.

Let me note that I think Hansson is right that some dyadic notion is essential to understanding supererogatory oversubscription, since it seems that it is more or less analytic that such cases involve satisfying an obligation by bringing about something else that is beyond the call. Here instead of Hansson's provisionally offered entailment relation, a dyadic agency notion is useful: $\mathbf{B}\mathrm{A}'_{\phi}\psi$ (Jane Doe brings it about that ψ by bringing it about that ϕ).[125] Then an alternative analysis not of BC but of the wide and important class Hansson places in formal focus, *oversubscription*, might be analyzed via a monadic notion like that in DWE or a similar system:

[125] See [McNamara, 2019].

$$\mathbf{O}\mathrm{BBA}\psi \;\&\; \mathbf{B}\mathrm{CBA}'_{\phi}\psi,$$

That is, it is obligatory for Jane that she bring about ψ and it is beyond the call for her to do so by bringing about ϕ. Here bring about ϕ is then an oversubscription to her obligation to bring about ψ.

Whether or not oversubscription is too narrow to serve as an analysis of supererogation, it is clearly an important species of supererogation, well worth scrutiny, and the proposal is interesting and suggestive, providing the first (to my knowledge) formal sketch of a framework for a highly significant species of supererogation.

7 Conclusion

In this essay, I have first of all tried to acquaint the reader with some of the key concepts in the conceptual neighborhood of supererogation (e.g. what one must do, what one ought to do, the least one can do, what is optional to do, what is a matter of indifference to do, what is good to do, bad to do, or neutral to do, what is permissibly suboptimal to do, what is beyond the call to do, what is praiseworthy to do, what is blameworthy to do). This was joined by exploring how this domain might be represented in a cohesive way, both by making sure that distinct concepts are represented distinctly, and by exploring various logical interrelationships that members of this rich family of concepts might bear to one another. Supererogation serves as a very fruitful focal point for exploring the logic of its family of associated notions. Some things seem clear. *There is a* richer and more expansive family of concepts than in the traditional scheme, and trying to cohesively model it is clearly a legitimate task for deontic logic. Whatever nuanced subtleties and refinements are yet to be made, there is a subject matter here, and there is a substantial tradition of conceptual work in this area that we can have learned from. However, it is also true that a lot of the exploration has tended to be in philosophical literature free of formal aims, and in forays into formal approaches to the area that have tended to be somewhat elementary, syntactic, underdeveloped, and mostly devoid of semantic underpinnings. In fairness, this is true of much of the early work in the 50s and 60s generally in deontic logic, even as formal semantics for intensional notions was emerging in corners that would soon become central, with Kanger, Kripke, Hintikka, (Bengst) Hansson and others.[126] As a result, a good dose of the work in this essay has been

[126] For a nice overview with an eye on deontic logic, see [Wolenski, 1990].

to draw out and round out some of the prior work that has been done. There is much that I had to leave out, even of work in that vein. Of the more formal work, there is not a lot. McNamara and Åqvist stand out for presenting syntactic, semantic and proof-theoretic presentations of some of the key target concepts in this area, along with soundness and completeness theorems.

It is also an underdeveloped area in that various other formal approaches surely make sense. For example, it surely makes sense to introduce agency operators explicitly and explore more robust integrations of agency with systems already developed (like McNamara's *DWE* framework or Åqvist's *Systematic Frame Constants* approach), and similarly for integrations of agency operators with new approaches aimed at representing some of the key concepts we've been trying to model in this essay. The use of dynamic logic, and explicit representations of actions, in the context of this family of concepts is clearly warranted as well. More explicit object language representations of a comparative operator and with that of exceeding the minimum or of being exceeded by the maximum, needs further exploration, as our reflections on McNamara's and Hansson's work indicate; likewise for doing more good than one has to do. As we noted for the more formal work mentioned (McNamara's and Åqvist's), the context is fairly classical, and so an adaptation of some of that work in a less classical context (.e.g. without inheritance) while still aiming at the considerable increases in expressive power makes good sense.[127] It also makes sense to explore modeling such a richer array of concepts from a contrastivist perspective[128], and likewise from a contextualist perspective.[129]. Trying to adapt such approaches to a more challenging and richer array of concepts than that of the traditional definitional scheme provides a good test case for such approaches, as well as promising to shed new light on the area, and perhaps revealing more distinctive features of moral and deontic discourse.

I hope this essay provides enough of a sense of prior work and concep-

[127]Some of this author's own work in that vein was excluded in the interest of space.

[128]See [Snedegar, Retrieved July 10 2019] for a general overview of contrastivism in ethics and references there, For an important formal contrastivist application, see [Cariani, 2013]. and for an application of contrastivism that preserves some of the connections between *ought*, *must*, and *may* that [McNamara, 1994] argued for, see [Snedegar, 2012],

[129]There is an industry rightfully inspired by Kratzer's groundbreaking contextualist approach to modals, and the first two chapters/papers in [Kratzer, 2012] provide a nice overview. For a defense against some recent challenges, see [Dowell, 2013]. See also [Portner, 2009], especially Chapters 3.1 and 5.2, for an accessible overview of Kratzer's contextualism.

tualizations of supererogation and its rich family of associated concepts to invite others to join in. Lastly, I hope it serves to support an aim I hold dear: that deontic logic has something to offer ethical theory.

References

[Anderson, 1956] Alan Ross Anderson. The formal analysis of normative systems. In Nicholas Rescher, editor, *The Logic of Decision and Action*, pages 147–213. University of Pittsburgh Press, Pittsburgh, 1956.

[Åqvist, 1985] Lennart Åqvist. A game-theoretical companion to Chisholm's ethics of requirement. *Acta Philosophica Fennica*, 38:327–347, 1985.

[Åqvist, 1990] L. Åqvist. Logical analysis of epistemic modality: An explication of the Bolding-Ekelöf degrees of evidential strength. In H.T. Klami, editor, *Rätt och sanning. Ett bevisteoretiskt symposium i Uppsala 26-27 maj 1989*, pages 43–54. lustus Förlag, Uppsala, 1990.

[Åqvist, 1993a] Lennart Åqvist. A completeness theorem in deontic logic with systematic frame constants. *Logique et Analyse*, 36:177–192, 1993.

[Åqvist, 1993b] Lennart Åqvist. Prima facie vs. toti-resultant obligations in deontic tense logic: Towards a formal reconstruction of the Richard Price-W. D. Ross theory. In C. Ciampi et al, editor, *Verso un sistema esperto giuridico integrale (Atti del Convegno celebrativo del venticinquennale dell'istituto per la Documentazione Giuridica del Consiglio Nationale delle Richerche, Firence, 1 dicembre 1993), 1995.* Padova: Cedam, 1993.

[Åqvist, 1997a] L. Åqvist. Systematic frame constants in defeasible deontic logic: A new form of Andersonian reduction. In D. Nute, editor, *Defeasible Deontic Logic: Essays in Nonmonotonic Normative Reasoning*, pages 59–77. Kluwer, Dordrecht/Boston/London, 1997.

[Åqvist, 1997b] Lennart Åqvist. Prima facie oughtness vs. oughtness all things considered in deontic logic: A Chisholmian approach. In E. Ejerhed and S Lindström, editors, *Logic, Action and Cognition. Trends in Logic (Studia Logica Library), vol. 2*, pages 89–96. Springer, Dordrecht, 1997.

[Åqvist, 1998] Lennart Åqvist. Prima facie obligations in deontic logic: A Chisholmian analysis based on normative preference structures. In Christoph Fehige, editor, *Preferences*, pages 135–155. de Gruyter, Hawthorne, 1998.

[Åqvist, 1999] L. Åqvist. Supererogation and offence in deontic logic: An analysis within systems of alethic modal logic with levels of perfection. In Rysiek Sliwinski, editor, *Philosophical Crumbs: Essays Dedicated to Ann-Mari Henschen-Dahlquist on the Occasion of Her Seventy-Fifth Birthday*, volume 49 of *Uppsala Philosophical Studies*, pages 261–276. Department of Philosophy, Uppsala University, Uppsala, Sweden, 1999.

[Åqvist, 2000] Lennart Åqvist. Three characterizability problems in deontic logic. *Nordic Journal of Philosophical Logic*, 5(2):65–82, 2000.

[Åqvist, 2002] Lennart Åqvist. Deontic logic. In Dov Gabbay and F. Guenthner, editors, *Handbook of Philosophical Logic, 2nd Edition*, volume 8, pages 147–264. Kluwer Academic Publishers, Dordrecht, 2nd edition, 2002.

[Arlo Costa, 2003] Horacio L. Arlo Costa. Review of Sven Ove Hansson, *The Structure of Values and Norms*, Cambridge University Press, 2001. *History and Philosophy of Logic*, 24:135–140, 2003.

[Belnap *et al.*, 2001] N Belnap, M Perloff, and M Xu. *Facing the Future: Agents and Choices in Our Indeterminist World.* Oxford University Press, New York, 2001.

[Belzer and Loewer, 1997] Marvin Belzer and Barry Loewer. Deontic logics of defeasibility. In D. Nute, editor, *Defeasible Deontic Logic*, Synthese Library, pages 45–57. Springer, Dordrecht, 1997.

[Cariani, 2013] Fabrizio Cariani. Ought and resolution semantics. *Noûs*, 47(3):534–558, 2013.

[Chellas, 1980] Brian F. Chellas. *Modal Logic: An Introduction.* Cambridge University Press, Cambridge, 1980.

[Chisholm and Sosa, 1966a] Roderick M. Chisholm and Ernest Sosa. Intrinsic preferability and the problem of supererogation. *Synthese*, 16:321–331, 1966.

[Chisholm and Sosa, 1966b] Roderick M. Chisholm and Ernest Sosa. On the logic of "intrinsically better". *American Philosophical Quarterly*, 3:244–249, 1966.

[Chisholm, 1963a] Roderick M. Chisholm. Contrary-to-duty imperatives and deontic logic. *Analysis*, 24:33–36, 1963.

[Chisholm, 1963b] Roderick M. Chisholm. Supererogation and offence: A conceptual scheme for ethics. *Ratio*, 5:1–14, 1963.

[Chisholm, 1964] Roderick M. Chisholm. The ethics of requirement. *American Philosophical Quarterly*, 1:147–153, 1964.

[Chisholm, 1974] Roderick M. Chisholm. Practical reason and the logic of requirement. In Stephan Korner, editor, *Practical Reason*, pages 1–17. Yale University Press, New Haven, 1974.

[Dowell, 2013] Janice Dowell. Flexible contextualism about deontic modals. *Inquiry*, 56:149–178, 2013.

[Driver, 1992] Julia Driver. The suberogatory. *Australasian Journal of Philosophy*, 70(3):286–295, 1992.

[Elgesem, 1993] Dag Elgesem. *Action Theory and Modal Logic.* Thesis, University of Oslo, 1993.

[Elgesem, 1997] Dag Elgesem. The modal logic of agency. *Nordic Journal of Philosophical Logic*, 2(2):1–46, 1997.

[Ewing, 1953] A. C. Ewing. *Ethics.* The Free Press, New York, 1953.

[Feinberg, 1961] Joel Feinberg. Supererogation and rules. *Ethics*, 71:276–288, 1961.

[Feldman, 1978] Fred Feldman. *Introductory Ethics.* Prentice-Hall, Inc., Englewood Cliffs, 1978.

[Forrester, 1975] Mary Forrester. Some remarks on obligation, permission, and supererogation. *Ethics*, 85:219–226, 1975.

[Goble, 1989] Lou Goble. A logic of better. *Logique et Analyse*, 32(27):297–318, 1989.

[Goble, 1990a] Lou. Goble. A logic of good, should, and would: Part 1. *Journal of Philosophical Logic*, 19(2):169–199, 1990.

[Goble, 1990b] Lou Goble. A logic of good, should, and would: Part 2. *Journal of Philosophical Logic*, 19(3):253–276, 1990.

[Governatori and Rotolo, 2005] Guido Governatori and Antonino Rotolo. On the axiomatisation of Elgesem's logic of agency and ability. *Journal of Philosophical Logic*, 34(4):403–431, 2005.

[Hansson, 1990a] Sven Ove Hansson. Defining "good" and "bad" in terms of "better". *Notre Dame Journal of Formal Logic*, 31(1):136–149, 1990.

[Hansson, 1990b] Sven Ove Hansson. Preference-based deontic logic (PDL). *Journal of Philosophical Logic*, 19(1):75–93, 1990.

[Hansson, 2001] Sven Ove Hansson. *The Structure of Values and Norms*. Cambridge University Press, Cambridge, 2001.

[Hansson, 2004] Sven Ove Hansson. Semantics for more plausible deontic logics. *Journal of Applied Logic*, 2:3–18, 2004.

[Hansson, 2006] Sven Ove Hansson. Ideal worlds – wishful thinking in deontic logic. *Studia Logica*, 82:329–336, 2006.

[Hansson, 2012] Sven Ove Hansson. Alternative semantics for deontic logic. In Dov Gabbay, John Horty, Xavier Parent, Ron van der Meyden, and Leon van der Torre, editors, *Handbook of Deontic Logic and Normative Systems, Volume 1*, page 445–497. College Publications, London, 2012.

[Hansson, 2013] Sven Ove Hansson. Representing supererogation. *Journal of Logic and Computation*, 25(2):443–451, 2013.

[Heyd, 1982] David Heyd. *Supererogation: Its Status in Ethical Theory*. Cambridge Studies in Philosophy. Cambridge University Press, Cambridge, 1982.

[Heyd, 2019] David Heyd. Supererogation. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, fall 2019 edition, 2019.

[Hilpinen and McNamara, 2013] Risto Hilpinen and Paul McNamara. Deontic logic: A historical survey and introduction. In Dov Gabbay, John Horty, Xavier Parent, Ron van der Meyden, and Leon van der Torre, editors, *Handbook of Deontic Logic and Normative Systems, Volume 1*, chapter 1, pages 1–134. College Publications, London, 2013.

[Horty, 2001] John F. Horty. *Agency and Deontic Logic*. Oxford University Press, Oxford England ; New York, 2001.

[Horty, 2002] John F. Horty. Review of the structure of values and norms, Sven Ove Hansson, Cambridge University Press, 2001. *Notre Dame Philosophical Reviews*, June, 2002.

[Hrushka and Joerden, 1987] Joachim Hrushka and Jan C Joerden. Su-

pererogation: vom deontologischen sechseck zum deontologischen zehneck. *Archiv fu Rechts und Sozialphilosophie*, 73:93–123., 1987.

[Humberstone, 1974] I. L. Humberstone. Logic for saints and heroes. *Ratio*, 16:103–114, 1974.

[Hurka, 1990] Thomas Hurka. Two kinds of satisficing. *Philosophical Studies*, 59:107–111, 1990.

[Joerden, 1991] Jan C Joerden. Supererogation. In H Burkhardt and B Smith, editors, *Handbook of Metaphysics and Ontology*. 1991.

[Joerden, 1998] Jan C. Joerden. On the logic of supererogation. *Jahrbuch für Recht und Ethik / Annual Review of Law and Ethics*, 6:145–159, 1998.

[Joerden, 2010] J.C. Joerden. *Logik im Recht*. Springer, Berlin/Heidelberg, 2nd edition, 2010.

[Joerden, 2012] Jan C. Joerden. Deontological square, hexagon, and decagon: A deontic framework for supererogation. *Logica Universalis*, 6:201–216, 2012.

[Jones and Sergot, 1996] Andrew Jones and Marek Sergot. A formal characterization of instituionalized power. *IGPL*, 4(3):429–445, 1996.

[Kratzer, 2012] Angelika Kratzer. *Modals and Conditionals*. Oxford Studies in Theoretical Linguistics. Oxford University Press, Oxford, 2012.

[Krogh and Herrestad, 1996] Christen Krogh and Henning Herrestad. Getting personal some notes on the relationship between personal and impersonal obligation. In Mark A. Brown and José Carmo, editors, *Deontic Logic, Agency and Normative Systems*, pages 134–153. Springer, 1996.

[Loewer and Belzer, 1983] Barry Loewer and Marvin Belzer. Dyadic deontic detachment. *Synthese*, 54:295–318, 1983.

[Loewer and Belzer, 1991] Barry Loewer and Marvin Belzer. "Prima facie" obligation. In Ernest Lepore, editor, *John Searle and His Critics*. Blackwell, Cambridge, 1991.

[Mares and McNamara, 1997] Edwin D. Mares and Paul McNamara. Supererogation in deontic logic: Metatheory for DWE and some close neighbours. *Studia Logica*, 59(3):397–415, 1997.

[McNamara, 1988] Paul McNamara. Supererogation, Act Utilitarianism and Urmson's Constraint: A Compatibility Thesis. In *Colloquium*. University of North Carolina, Greensboro, 1988.

[McNamara, 1990] Paul McNamara. *The Deontic Quaddecagon*. Dissertation, University of Massachusetts, 1990.

[McNamara, 1993] Paul McNamara. Doing Well Enough: Toward a Logic for Commonsense Morality. In Andrew Jones and M. Sergot, editors, *DEON '94: Second International Workshop on Deontic Logic in Computer Science*, pages 165–197. Tano, Norway, Oslo, 1993.

[McNamara, 1994] Paul McNamara. Must I do what I ought? (typescript). 1994.

[McNamara, 1996a] Paul McNamara. Doing well enough: Toward a logic for commonsense morality. *Studia Logica*, 57(1):167–192, 1996.

[McNamara, 1996b] Paul McNamara. Making room for going beyond the call. *Mind*, 105(419):415–450, 1996.

[McNamara, 1996c] Paul McNamara. Must I do what I ought? (or will the least I can do do?). In Mark A. Brown and José Carmo, editors, *Deontic Logic, Agency and Normative Systems*, pages 154–173. Springer Verlag, New York, 1996.

[McNamara, 1999] Paul McNamara. Doing well enough in an Andersonian-Kangerian framework. In Paul McNamara and Henry Prakken, editors, *Norms, Logics and Information Systems: New Studies in Deontic Logic and Computer Science*, pages 181–198. IOS Press, Washington, DC, 1999.

[McNamara, 2000] Paul McNamara. Toward a framework for agency, inevitability, praise and blame. *Nordic Journal of Philosophical Logic*, 5(2):135–160, 2000.

[McNamara, 2004] Paul McNamara. Agential obligation as non-agential personal obligation plus agency. *Journal of Applied Logic*, 2(1):117–152, 2004.

[McNamara, 2006] Paul McNamara. Supererogation and utilitarianism revisited: An inroad to the structure of common sense morality. *Colloquium*. University of London, Institute of Philosophy, School of Advanced Studies, 2006.

[McNamara, 2011a] Paul McNamara. Praise, blame, obligation, and DWE: Toward a framework for the classical conception of supererogation and kin. *Journal of Applied Logic*, 9:153–170, 2011.

[McNamara, 2011b] Paul McNamara. Supererogation, inside and out: Toward an adequate scheme for common sense morality. In Mark Timmons, editor, *Oxford Studies in Normative Ethics, Volume I*, pages 202–235. Oxford University Press, Oxford, 2011.

[McNamara, 2019] Paul McNamara. Toward a systematization of logics for monadic and dyadic agency & ability, revisited. *Filosofiska Notiser*, 6:157–188, 2019.

[McNamara, Forthcoming] Paul McNamara. A Natural Conditionalization of the DWE Framework. In *Agency, Normative Systems, Artifacts, and Beliefs: Essays in Honor of Risto Hilpinen*. Paul McNamara, Andrew Jones, and Mark Brown (eds.). Forthcoming.

[Meinong, 1894] Alexius Meinong. Psychologisch ethische untersuchungen zur werttheorie. In R. Haller and R. Kindinger, editors, *Alexius Meinong Gesamtausgabe, Vol. III, Abhandlungen zur Werttheorie*, pages 3–94. Akademische Druck- u. Velagsanstalt, Graz, 1894.

[Meinong, 1968] Alexius Meinong. Ethische bausteine (nacgelassenes fragment). In R. Haller and R. Kindinger, editors, *Alexius Meinong Gesamtausgabe Vol III Abhandlungen zur Werttheorie*, pages 659–724. Akademische Druck- und Verlagsanstalt, Graz, 1968.

[Mellema, 1987] Gregory Mellema. Quasi-supererogation. *Philosophical Studies*, 52:141–150, 1987.

[Mellema, 1991] Gregory Mellema. *Beyond the Call of Duty: Supererogation,*

Obligation, and Offence. SUNY Pr, Albany, 1991.

[Moore, 1903] G. E. Moore. *Principia Ethica*. Cambridge University Press, Cambridge, 1903.

[Moore, 1912 1965 edition] G. E. Moore. *Ethics*. Oxford University Press, New York, 1912 (1965 edition).

[Moretti, 2004] Alessio Moretti. Geometry for modalities? Yes: Through n-opposition theory. In J.Y. Beziau, A. Costa-Leite, and A. Facchini, editors, *Universal Logic*, pages 102–145. 2004.

[Moretti, 2009] Alessio Moretti. *The Geometry of Logical Opposition*. Thesis, University of Neuchâtel, 2009.

[Portner, 2009] Paul Portner. *Modality*. Oxford Surveys in Semantics and Pragmatics. Oxford University Press, Oxford, 2009.

[Prior, 1962 1955] A. N. Prior. *Formal Logic*. Oxford University Press, Oxford, second edition, 1962 [1955].

[Purtill, 1973] Richard Purtill. Meinongian deontic logic. *Philosophical Forum (Boston)*, 4:585–592, 1973.

[Schotch and Jennings, 1981] Peter K. Schotch and Raymond E. Jennings. Non-kripkean deontic logic. In Risto Hilpinen, editor, *New Studies in Deontic Logic*, pages 149–162. Reidel, Dordrecht, 1981.

[Sergot, 2013] Marek Sergot. Normative positions. In Dov Gabbay, John Horty, Xavier Parent, Ron van der Meyden, and Leon van der Torre, editors, *Handbook of Deontic Logic and Normative Systems*, volume 1, pages 353–406. College Publications, London, 2013.

[Snedegar, 2012] Justin Snedegar. Contrastive semantics for deontic modals. In M. Blaauw, editor, *Contrastivism in Philosophy*, pages 116–133. Routledge, Abingdon, 2012.

[Snedegar, Retrieved July 10 2019] Justin Snedegar. *Ethics and contrastivism*. Internet Encyclopedia of Philosophy, Retrieved July 10, 2019.

[Urmson, 1958] J.O. Urmson. Saints and heroes. In A. I. Melden, editor, *Essays in Moral Philosophy*, pages 198–216. University of Washington Press, Seattle, 1958.

[von Wright, 1951] G. H. von Wright. Deontic logic. *Mind*, 60:1–15, 1951.

[von Wright, 1953] G. H. von Wright. *An Essay in Modal Logic*. Humanities Press, New York, 1953.

[von Wright, 1963] G. H. von Wright. *Norm and Action: A Logical Enquiry*. Humanities Press, New York, 1963.

[Wessels, 2002] Ulla Wessels. *Die gute Samariterin. Zur Struktur der Supererogation*. de Gruyter, Berlin, 2002.

[Wessels, 2003] Ulla Wessels. Die gute samariterin: Zur struktur der supererogation (ideen und argumente). *Tijdschrift voor Filosofie*, pages 65(4) 776–777, 2003.

[Wessels, 2015] Ulla Wessels. Beyond the call of duty: The structure of a moral region. In Christopher Crowley, editor, *Supererogation*, pages 87–104.

Cambridge University Press, Cambridge, 2015.

[Wolenski, 1990] Jan Wolenski. Deontic logic and possible world semantics: A historical sketch. *Studia Logica*, pages 273–282, 1990.

[Zimmerman, 1996] Michael J. Zimmerman. *The Concept of Moral Obligation*. Cambridge University Press, Cambridge, 1996.

Paul McNamara
University of New Hampshire
Email: paulm@unh.edu

4
Deontic Logic and Changing Preferences

Johan van Benthem and Fenrong Liu

Abstract. The normative realm involves deontic notions such as obligation or permission, as well as information about relevant actions and states of the world. This mixture is not static, given once and for all. Both information and normative evaluation available to agents are subject to changes with various triggers, such as learning new facts or accepting new laws. This chapter explores models for this setting in terms of dynamic logics for information-driven agency. Our paradigm will be dynamic-epistemic logics for knowledge and belief, and their current extensions to the statics and dynamics of agents' preferences. Here the link with deontics is that moral reasoning may be viewed as involving preferences of the acting agent as well as preferences of moral authorities such as lawgivers, one's conscience, or yet others. In our presentation of preference based agency, we discuss a large number of themes: primitive 'betterness' order versus reason-based preferences (employing a model of 'priority graphs'), the entanglement of preference and informational attitudes such as belief, interactive social agents, and scenarios with long-term patterns emerging over time. Specific deontic issues considered include paradoxes of deontic reasoning, acts of changing obligations, and changing norm systems. We conclude with some further directions, such as multi-agency and games, plus pointers to related work, including different paradigms for looking at these same phenomena.

1 Agency, information, and preference **309**

2 Dynamic logics of knowledge and belief change **310**
2.1 Epistemic logic and semantic information 310
2.2 Dynamic logic of public announcement 312
2.3 From knowledge to belief 314
2.4 Dynamic logics of belief change 317
2.5 General dynamic methodology and its applications 319

Handbook of Deontic Logic and Normative Systems, Volume 2

3 Deontic logic as preference logic **320**

4 Static preference logic **322**
4.1 General modal preference logic 322
4.2 Special features of preference 324

5 World based dynamics of preference change **325**
5.1 Betterness change 326
5.2 Deriving changes in defined preferences 327
5.3 General formats for betterness change 327

6 Reason-based preferences **329**
6.1 Priority based preference 329
6.2 Pre-orders 330
6.3 Static logic and representation theorem 330
6.4 Priority dynamics and graph algebra 331

7 A two-level view of preference **332**
7.1 Harmony of world order and reasons 332
7.2 Correlated dynamics 333
7.3 Additional dynamics: language change 334

8 Combining evaluation and information **334**
8.1 Generic preference with knowledge 335
8.2 Generic preference with belief 336
8.3 Other entanglements of preference and normality 336
8.4 Preference change and belief revision 337

9 Deontic reasoning, changing norms and obligations **338**
9.1 Triggers for deontic actions and events 338
9.2 Unary and dyadic obligation on ordering models 339
9.3 Reasons and dynamics in classical deontic scenarios 340
9.4 Typology of change at two levels 342
9.5 Norm change 343
9.6 Entangled changes 343

10 Further directions **345**
10.1 Language, speech acts, and agency 345
10.2 Multi-agency and groups 345
10.3 Games and dependent behavior 346
10.4 Temporal perspective 347

10.5 Fine-structure of information 348
10.6 Digression: numerical strength 349
10.7 Probability 350

11 Appendix: relevant strands in the literature 350

12 Conclusion 353

1 Agency, information, and preference

Agents pursue goals in this world, acting within constraints in terms of their information about what is true, as well as norms about what is right. The former dimension typically involves acts of inference, observation, as well as communication and other forms of social interaction. The latter dimension involves evaluation of situations and actions, 'coloring' the agents' view of the world, and driving their desires, decisions, and actions in it. A purely informational agent may be rational in the sense of clever reasoning, but a *reasonable* agent is one whose actions are in harmony with what she wants. The two dimensions are intimately related. For instance, what we want is influenced by what we believe to be true as well as what we prefer, and normally also, we only seek information to further goals that we desire.

This balance of information and evaluation is not achieved once and for all. Agents must constantly cope with new information, either because they learn more about the current situation, or because the world has changed. But equally well, agents constantly undergo changes in evaluation, sometimes by intrinsic changes of heart, but most often through events with normative impact, such as accepting a command from an authority. These two forms of dynamics, too, are often entangled: for instance, learning more about the facts can change my evaluation of a situation.

A third major aspect of agency is its social interactive character. Even pure information flow is often driven by an epistemic gradient: the fact that different agents know different things leads us to communicate, whether in cooperative inquiry or adversarial argumentation, perhaps until a state of equilibrium is reached such as common knowledge or common belief. But also more complex forms of interaction occur, such as merging beliefs in groups of agents, where differences in informational authority may play a crucial role. Again, very similar phenomena play on the normative side. Norms, commitments and duties usually involve

other agents, both as their source and as their target, and whole institutions and societies are constructed in terms of social choice, shared norms and rules of behavior.

In this chapter, we take current "dynamic-epistemic logics" as our model for the above phenomena, informational and preferential, and we show how this perspective transfers to normative reasoning and deontic logic. We will highlight two main themes: (a) making the dynamic *actions and events* that drive real deontic scenarios, such as commands or permissions, an explicit part of our analysis, and (b) exploring more finely structured *reasons for deontic preferences*. Important side-themes linked with these are (c) the entanglement of obligation, information, knowledge and belief, and (d) the importance of multi-agent scenarios, such as games, in the deontic realm. Our treatment will be brief, and for a more elaborate sample of this style of thinking about the normative realm, we refer to [Benthem *et al.*, 2014].

In pursuing the specific technical paradigm of this chapter, we are not at all denying the existence of other valid approaches to deontic dynamics or further themes covered, and we will provide a number of references to other relevant strands in the literature.

2 Dynamic logics of knowledge and belief change

Before analyzing preference or related deontic notions, we first develop the basic methodology of this chapter for the purely informational case, where the first dynamic-epistemic logics arose in the study of information update and information exchange between agents.

2.1 Epistemic logic and semantic information

Dynamic logics of agency need an account of underlying static states that can be modified by suitable triggers: actions or events. Such states usually come from existing systems in philosophical or computational logic whose models can serve as static snapshots of the dynamic process.

In this chapter, we start with a traditional modal base system of epistemic logic, referring to the standard literature for details (cf. [Fagin *et al.*, 1995] and [Blackburn *et al.*, 2001]).

Definition 1. *Let a set of propositional variables Φ be given, as well as a set of agents A. The epistemic language is defined by the syntax rule*

$$\varphi := \top \mid p \mid \neg\varphi \mid \varphi \wedge \psi \mid K_a\varphi \qquad \textit{where } p \in \Phi,\ a \in A.$$

Remark: Single agents, interacting agents, and groups. For convenience, we will focus on single agents in this chapter – something that still allows us to describe interacting individual agents where needed through iterations of modalities. Epistemically important notions with groups themselves as agents, such as 'common knowledge' or 'distributed knowledge', are deferred to our discussion at the end. Group actors are also important in the deontic realm, involving collective commitments or duties, but we will only touch upon this theme occasionally.

Semantic models for the epistemic language encode agents' 'information ranges' in the form of equivalence classes of binary uncertainty relations for each agent.[1] These models support a standard compositional truth definition.

Definition 2. *An* epistemic model *is a tuple* $\mathfrak{M} = (W, \{\sim_a\}_{a\in A}, V)$ *with* W *a set of epistemically possible states (or 'worlds'),* $\sim_a$ *an equivalence relation on* W*, and* V *a valuation function from* Φ *to subsets of* W*.*

Definition 3. *For an epistemic model* $\mathfrak{M} = (W, \{\sim_a \mid a \in A\}, V)$ *and any world* $s \in S$*, we define* $\mathfrak{M}, s \models \varphi$ *(epistemic formula* φ is true in $\mathfrak{M}$ at s*) by induction on the structure of the formula* φ*:*

1. $\mathfrak{M}, s \models \top$ *always.*
2. $\mathfrak{M}, s \models p$ iff $s \in V(p)$.
3. $\mathfrak{M}, s \models \neg\varphi$ iff *not* $\mathfrak{M}, s \models \varphi$.
4. $\mathfrak{M}, s \models \varphi \wedge \psi$ iff $\mathfrak{M}, s \models \varphi$ *and* $\mathfrak{M}, s \models \psi$.
5. $\mathfrak{M}, s \models K_a\varphi$ iff *for all* t *with* $s \sim_a t : \mathfrak{M}, t \models \varphi$.

Using equivalence relations in our models yields the well-known modal system **S5** for each individual knowledge modality, without interaction laws for different agents. Just for concreteness, we state this basic fact here:

Theorem 4. *Basic epistemic logic is axiomatized completely by the axioms and inference rules of the modal system* ***S5*** *for each separate agent.*

[1] The approach of this chapter will also work on models with more general relations such as transitive and reflexive pre-orders, but we start with this easily visualizable epistemic case for expository purposes.

Few researchers see our basic modalities and the simple axioms of modal **S5** as expressing genuine properties of 'knowledge' – thus making earlier polemical discussions of epistemic 'omniscience' or 'introspection' expressed by these axioms obsolete. Our interpretation of the above notions is as describing the *semantic information* that agents have available (cf. [Benthem, 2014]), being a modest but useful building block in analyzing more complex epistemic and deontic notions. We will allow ourselves the use of the word 'know' occasionally, however: old habits die hard. [2]

Now comes our first key theme. Static epistemic logic describes what agents know on the basis of their current semantic information. But information flows, and a richer story must also include dynamics of actions that produce and modify information. We now turn to the simplest case of this dynamics: reliable public announcements or public observations, that shrink the current information range.

2.2 Dynamic logic of public announcement

The pilot for the methodology of this paper is 'public announcement logic' (PAL), a toy system describing a combination of epistemic logic and one dynamic event, namely, *announcement* of new 'hard information' expressed in some proposition φ that is true at the actual world. The corresponding 'update action' $!\varphi$ transforms a current epistemic model $\mathfrak{M}, s$ into its definable submodel $\mathfrak{M}|\varphi, s$ where all worlds that did not satisfy φ have been eliminated. This model update is the basic scenario of obtaining information in the realm of science but also of common sense, by shrinking one's current epistemic range of uncertainty.[3]

To describe this phenomenon, the *language* of PAL has two syntactic levels working together, using formulas for propositions as well as action expressions for announcements:

$$\varphi := \top \mid p \mid \neg\varphi \mid \varphi \wedge \psi \mid K_a\varphi \mid [A]\varphi$$
$$A := !\varphi$$

In particular, the new dynamic formula $[!\varphi]\psi$ says that "after updating with the true proposition φ, formula ψ holds":

[2] There is a fast-growing literature on more sophisticated logical analyses of genuine knowledge (cf. [Holliday, 2012], [Benthem and Pacuit, 2011], [Shi, 2014]), which also seems relevant to modeling and reasoning in the deontic realm. However, the main points to be made in this chapter are orthogonal to these additional refinements.

[3] The name 'public announcement logic' may be unfortunate, since the logic PAL describes updates with hard information from whatever source, but no consensus has emerged yet on a rebaptism.

$\mathfrak{M}, s \models [!\varphi]\psi$ iff if $\mathfrak{M}, s \models \varphi$, then $\mathfrak{M}|\varphi, s \models \psi$.

This language can make characteristic assertions about knowledge change such as $[!\varphi]K_a\psi$, which states what agent a will know after having received the hard information that φ. In particular, the knowledge change before and after an update can be captured by so-called *recursion axioms*, a sort of recursion equations for the 'dynamical system' of PAL, relating new knowledge to knowledge that agents had before. Here is the complete logical system for information flow under public announcement (two original sources are [Gerbrandy, 1999], [Plaza, 1989]):

Theorem 5. *PAL is axiomatized completely by the usual laws of the static epistemic base logic plus the following recursion axioms:*

1. $[!\varphi]q \leftrightarrow (\varphi \rightarrow q)$ *for atomic facts q*
2. $[!\varphi]\neg\psi \leftrightarrow (\varphi \rightarrow \neg[!\varphi]\psi)$
3. $[!\varphi](\psi \wedge \chi) \leftrightarrow ([!\varphi]\psi \wedge [!\varphi]\chi)$
4. $[!\varphi]K_a\psi \leftrightarrow (\varphi \rightarrow K_a[!\varphi]\psi)$

These elegant principles analyze reasoning about epistemic effects of receiving hard information, through observation, communication, or other reliable means. In particular, the knowledge law reduces knowledge after new information to 'conditional knowledge' that the agent had before, but in a subtle recursive manner. This prudence of design for PAL is necessary since the dynamic process of information update can typically change truth values of epistemic assertions over time. Perhaps, initially, I did not know that p, but after the event $!p$, I do.

There are several noteworthy features to this approach. We have already stressed the recursive nature of reducing new knowledge to preexisting knowledge, a feature that is typical of dynamical systems. Also, the precise way in which this happens involves breaking down, not the announced propositions (as one might expect), but the 'postconditions' behind the dynamic modalities $[!\varphi]$ compositionally on the basis of their syntactic shape.[4]

Next, as things stand here, repeating these steps, the stated features drive a 'reduction process' taking every formula of our dynamic-epistemic language eventually to an equivalent formula inside the static epistemic language. In terms of semantics and expressive power, this

[4]One can have compound informational actions, of course, but these would rather be modeled in an extended syntax of programs over atomic announcement actions.

means that a current static model 'pre-encodes' all information about what might happen when agents communicate what they know. Moreover, in terms of the logic, the reduction procedure means that PAL is *axiomatizable* and *decidable*, since it inherits these features from the epistemic base logic.

However, it is important to note that sweeping dynamics-to-statics reduction is not an inevitable feature of dynamic-epistemic analysis. In recent semantics for PAL, available sequences of updates are constrained by *protocols* that restrict available events in the current process of inquiry. In that case, no reduction is possible to the base logic, and the dynamic logic, though still employing recursion equations, and remaining axiomatizable and decidable, comes to encode a genuine new kind of 'procedural information' (cf. [Benthem *et al.*, 2009a]). Protocols also make sense for deontic purposes, because of the procedural character of much normative behavior, and we will briefly return to this perspective at the end of this chapter.

In what follows, PAL will serve as a pilot example for many other complex cases, for example, changes in beliefs, preferences, and obligations. In each case, the 'triggering events' can be different: for instance, beliefs can change by signals of different force: hard or more 'soft', and obligations can change through actions of commanding by a normative authority. In many cases, the domain of the model does not change, but rather its *ordering pattern*.[5] However, the general recursive methodology of PAL will remain in force, though in each case, with new twists.

2.3 From knowledge to belief

Knowledge rests on hard information, but most of the information that we have and act on is soft, giving rise to *beliefs*, that are not always true, and that can be revised when shown inadequate. One can think of learning from error as the more creative ability, beyond mere recording of reliable information in the agent's environment.

Again we need to start with a convenient static base for our investigation. One powerful model for soft information and belief reflects the intuition that we believe those things that hold in the *most plausible* worlds in our epistemic range. I believe that this train will take me home on time, even though I do not know that it will not suddenly fly

[5] One example of this approach, even in the epistemic realm, are 'link cutting' versions of updating after announcement: cf. [Liu, 2004], [Snyder, 2004], [Benthem and Liu, 2007]. Such transformations will be used later on in scenarios where we may want to return to worlds considered earlier in the process.

away from the tracks. But the worlds where it stays on track are more plausible than those where it flies off, and among the latter, those where it arrives on time are more plausible than those where it does not.

The long history for this way of modeling belief includes non-monotonic logic in artificial intelligence [Shoham, 1988; Boutilier, 1992; Lamarre and Shoham, 1994; Friedman and Halpern, 1997; Friedman and Halpern, 1999],[6] the semantics of natural language (cf. [Veltman, 1996]), as well as the philosophical literature on epistemology and games (cf. [Stalnaker, 1996; Baltag and Smets, 2008]). The common intuition of relative plausibility leads to the following semantics:

Definition 6. *An* epistemic-doxastic model $\mathfrak{M} = (W, \{\sim_a\}_{a\in A}, \{\leq_a\}_{a\in A}, V)$ *consists of an epistemic model* $(W, \{\sim_a\}_{a\in A}, V)$ *as before, while the additional relations* $\leq_a$ *are binary comparative plausibility pre-orders for agents between worlds.*

Intuitively, these comparison orders might well be *ternary* $\leq_{a,s} xy$ saying that, in world s, agent a considers world x at least as plausible as y.[7] For convenience in this chapter, however, our semantics assumes that plausibility orderings are the same for epistemically indistinguishable worlds: that is, agents know their plausibility judgements. Assuming that plausibility is a pre-order, i.e., reflexive and transitive, but not necessarily connected, leaves room for the existence of genuinely incomparable worlds – but what we have to say also holds for the special case of *connected* pre-orders where any two worlds are comparable.[8]

As with epistemic models, however, our style of logical analysis will work largely independently from specific design decisions about the ordering, important though these may be in specific applications.

One can interpret many languages in the above structures. In what follows, we work with modal formalisms for the usual reasons of perspicuous formulation and low complexity (cf. [Blackburn *et al.*, 2007]).

[6]Non-monotonic logic has been a continuing source of inspiration for logics of belief, preference, and even deontic logic, in the treatment of conditional belief or conditional obligation. As another type of illustration, the two last-mentioned papers also show analogies with our Section 6 on reasons for preference.

[7]In particular, ternary world-dependent plausibility relations are found in the semantics of conditional logic: cf. [Lewis, 1973; Spohn, 1988], models for games: cf. [Stalnaker, 1999; Benthem, 2014], as well as in recent logical analyses of major paradigms in epistemology: [Holliday, 2012].

[8]Connected orders are equivalent to the 'sphere models' of conditional logic or belief revision theory (cf. [Grove, 1988; Segerberg, 2001]) – but in these areas, too, a generalization to pre-orders has been proposed: for instance, in: [Burgess, 1984; Shoham, 1988] and [Veltman, 1985].

First, there is *absolute belief* as truth in all most plausible worlds:

$\mathfrak{M}, s \models B_a\varphi$ iff $\mathfrak{M}, t \models \varphi$ *for all those worlds* $t \sim_a s$ *that are maximal in the order* $\leq_a$ *xy in the* epistemic $\sim_a$*-equivalence class of the world* s.[9]

But the more general notion in our models is that of a *conditional belief*:

$\mathfrak{M}, s \models B_a^{\psi}\varphi$ iff $\mathfrak{M}, t \models \varphi$ *for all those worlds* $t \sim_a s$ *that are maximal for* $\leq_a$ *xy in the set* $\{u \mid s \sim_a u$ *and* $\mathfrak{M}, u \models \psi\}$.

Conditional beliefs generalize absolute beliefs, which are now definable as $B_a^{\top}\varphi$. They *pre-encode* absolute beliefs that we will have *if* we learn certain things. Indeed, the above semantics for $B_a^{\psi}\varphi$ is formally similar to that for conditional assertions $\psi \Rightarrow \varphi$. This allows us to use known results from [Burgess, 1984], [Veltman, 1985]:[10]

Theorem 7. *The logic of* $B_a^{\psi}\varphi$ *is axiomatized by standard propositional logic plus the laws of conditional logic over pre-orders.*

Deductively stronger modal logics also exist in this area, such as the popular system **KD45** for absolute belief. The structural content of additional axioms can be determined by standard modal frame correspondence techniques (see [Blackburn *et al.*, 2007; Benthem, 2010]).

Digression: Further relevant attitudes. Modeling agency with just the notions of knowledge and belief is mainly a tradition inherited from the literature. In a serious study of agency the question needs to be raised afresh what is our natural repertoire of attitudes triggered by information. As one interesting example, the following operator has emerged recently, in between knowledge and belief qua strength. Intuitively, 'safe belief' is belief that agents currently have which cannot be falsified by receiving true new information.[11] Over connected epistemic plausibility models $\mathfrak{M}$, its can be defined as follows:

[9] Maximality intuitions may fail in models with infinite sequences in the plausibility ordering. However, there, natural reformulations arise such as the following (see, e.g., [Girard, 2008]): $\mathfrak{M}, s \models B^{\psi}\varphi$ iff $\forall t \sim s : \exists u :$ ($t \preceq u$ *and* $\mathfrak{M}, u \models \psi$ *and* $\forall v \sim s$: (*if* $u \preceq v$ *and* $\mathfrak{M}, v \models \psi$, *then* $\mathfrak{M}, v \models \varphi$)).

[10] For some recent completeness theorems in deontic logic over Hanson-style betterness orders paralleling this line of work in conditional logic, see our final section on related literature in this chapter.

[11] This notion has been proposed independently in AI, cf. [Shoham and Leyton-Brown, 2008], philosophy, cf. [Stalnaker, 2006], learning theory, and game theory, cf. [Baltag *et al.*, 2011; Baltag *et al.*, 2009].

Definition 8. *The modality of* safe belief $B_a^+\varphi$ *is interpreted as follows:*

$$\mathfrak{M}, s \models B_a^+\varphi \quad \text{iff} \quad \textit{for all worlds } t \sim_a s\textit{: if } s \leq_a t\textit{, then } \mathfrak{M}, t \models \varphi.$$

Thus, the formula φ is to be true in all accessible worlds that are at least as plausible as the current one. This includes the most plausible worlds, but it need not include all epistemically accessible worlds, since the latter may have some less plausible than the current one. The logic for safe belief is just **S4**, since it is in fact the simplest modality over the plausibility order.

A notion like this has the conceptual advantage of making us see that agents can have more responses to information than just knowledge and belief.[12] But there is also the technical advantage that the simple modality of safe belief can define more complex notions such as conditional belief (see [Lamarre, 1991], [Boutilier, 1994], [Benthem, 2014]) – which can lead to simplifications of logics for agency.

2.4 Dynamic logics of belief change

Having set up the basic attitudes, we now want to deal with explicit acts or events that update not just knowledge, but also agents' beliefs.[13]

Hard information The first obvious triggering event are the earlier public announcements of new hard information. Their complete logic of belief change can be developed in analogy with the earlier dynamic epistemic logic PAL, again via world elimination. Its key recursion axiom for new beliefs uses conditional beliefs:

Fact 9. *The following formula is valid in our semantics:*

$$[!\varphi]B_a\psi \leftrightarrow (\varphi \rightarrow B_a^{\varphi}[!\varphi]\psi)$$

To keep the complete dynamic language in harmony, we then also need a recursion axiom for the conditional beliefs that are essential here:

Theorem 10. *The dynamic logic of conditional belief under public announcements is axiomatized completely by*

(a) any complete static modal logic for the model class chosen,

[12] Other relevant notions extending the usual epistemic-doxastic core vocabulary include the 'strong belief' of [Stalnaker, 2006], [Baltag and Smets, 2008].

[13] For a much more extensive up-to-date survey of logic-based belief revision, cf. [Benthem and Smets, 2015].

(b) the earlier PAL recursion axioms for atomic facts and for the Boolean operations,

(c) the following recursion axiom for conditional beliefs:

$$[!\varphi]B_a^{\chi}\psi \leftrightarrow (\varphi \rightarrow B_a^{\varphi \wedge [!\varphi]\chi}[!\varphi]\psi)$$

This analysis also extends to the further notion of safe belief, with the following even simpler recursion law:

Fact 11. *The following PAL-style axiom holds for safe belief:*

$$[!\varphi]B_a^{+}\psi \leftrightarrow (\varphi \rightarrow B_a^{+}(\varphi \rightarrow [!\varphi]\psi)).$$

Using this equivalence, which behaves more like the original central *PAL* axiom, one can show that safe belief has its intuitively intended features. Safe belief in factual propositions (i.e., those not containing epistemic or doxastic operators) remains safe belief after updates with hard factual information.[14]

Soft information But belief change also involves more interesting triggers, depending on the quality of the incoming information, or the trust agents place in it. 'Soft information upgrade' does not eliminate worlds as what hard information does, but it rather *changes the plausibility order*, promoting or demoting worlds according to their properties. Here is one widely used way in which this order change can happen: an act of 'radical', or 'lexicographic' upgrade.[15]

Definition 12. *A* radical upgrade $\Uparrow\varphi$ *changes the current plausibility order* $\leq$ *between worlds in* $\mathfrak{M}, s$ *to create a new model* $\mathfrak{M}\Uparrow\varphi, s$ *where all* φ*-worlds in* $\mathfrak{M}, s$ *become better than all* $\neg\varphi$*-worlds, while, within those two zones, the old plausibility order* $\leq$ *remains as it was.*

No worlds are eliminated here, it is the ordering pattern that adapts. There is a matching upgrade modality for this in our dynamic language:

$$\mathfrak{M}, s \models [\Uparrow\varphi]\psi \quad \text{iff} \quad \mathfrak{M}\Uparrow\varphi, s \models \psi.$$

This extended setting supports one more dynamic completeness theorem (cf. [Benthem, 2007]).

[14] Unlike with plain belief, the latter recursion does not involve a move to an irreducible new notion of 'conditional safe belief'. Indeed, given a definition of conditional belief in terms of safe belief, the more complex recursion law in Theorem 10 becomes derivable from the above simple principle.

[15] Henceforth, in this section, with a few exceptions, we will drop mention of epistemic accessibility, and focus on plausibility order only.

Theorem 13. *The logic of radical upgrade is axiomatized completely by*

(a) a complete axiom system for conditional belief on the static models,

(b) the following recursion axioms for postconditions:

$$
\begin{aligned}
&[\Uparrow\varphi]q \;\leftrightarrow\; q, \quad \textit{for all atomic proposition letters } q\\
&[\Uparrow\varphi]\neg\psi \;\leftrightarrow\; \neg[\Uparrow\varphi]\psi\\
&[\Uparrow\varphi](\psi\wedge\chi) \;\leftrightarrow\; ([\Uparrow\varphi]\psi\wedge[\Uparrow\varphi]\chi)\\
&[\Uparrow\varphi]B^{\chi}\psi \;\leftrightarrow\; (E(\varphi\wedge[\Uparrow\varphi]\chi)\wedge B^{\varphi\wedge[\Uparrow\varphi]\chi}[\Uparrow\varphi]\psi)\\
&\qquad\vee(\neg E(\varphi\wedge[\Uparrow\varphi]\chi)\wedge B^{[\Uparrow\varphi]\chi}[\Uparrow\varphi]\psi)
\end{aligned}
$$

Here the operator 'E' is the standard existential epistemic modality, and we need to add a simple recursion axiom for knowledge under upgrade, that we forego here.[16]

There are many further policies for changing plausibility order. For instance, 'conservative upgrade' $\uparrow\varphi$ only puts the *most plausible* φ-worlds on top in the new model, leaving the rest in their old positions. [Rott, 2006] is an excellent philosophical source for the variety of policies found in belief revision theory that is not tied to the specific dynamic logic methodology employed in this chapter. For general results on complete dynamic logics of belief change in our style, see [Benthem, 2007; Baltag and Smets, 2008] and [Benthem, 2011]. The most up-to-date survey as of now is the Handbook chapter [Benthem and Smets, 2015].

A plea for a little patience. Readers wondering why we introduce all these different logics for information, knowledge and belief, may want to think already about their counterparts for deontic notions. In fact, analogies are easy to find. For instance, concerning our static repertoire, safe belief is like the 'betterness' modality that we will use later to describe preference. And as for our dynamic repertoire, the distinction between hard and soft information has obvious counterparts in different forces that we can give to commands coming from moral authorities.

2.5 General dynamic methodology and its applications

We have spent quite some time on the above matters because they represent a general methodology of *model transformation* that works for many further phenomena, including changes in preference, and the even richer deontic scenarios that we will be interested in eventually.

[16] As before, it is easy to extend this dynamic analysis of soft upgrade to simpler recursion axioms for the case of safe belief.

Model transformations of relevance to agency can be much more drastic than what we have seen here, extending the domains of available worlds and modifying their relational structure accordingly. In the dynamic epistemic logic of general observation DEL, different agents can have different access to the current informational event, as happens in card games, communication with security restrictions, or other social scenarios. This requires generalizing PAL as well as the above logics of belief change, using a mechanism of 'product update' to create more complex new models whose size can even increase (cf. [Baltag *et al.*, 1998; van Ditmarsch *et al.*, 2007; Benthem, 2011]).

Appropriately extended update mechanisms have been applied to many further aspects of agency: changes in intentions [Roy, 2009, Icard III], trust [Holliday, 2009], inference [Velazquez-Quesada, 2009], questions and inquiry ([Benthem and Minica, 2009]), as well as complex scenarios in games [Otterloo, 2005; Benthem, 2014] and social information phenomena generally [Seligman *et al.*, 2013; Baltag *et al.*, 2013; Liu *et al.*, 2014; Hansen and Hendricks, 2014]. There are also studies tying update mechanisms to more general dynamic logics of graph change, such as [Aucher *et al.*, 2009a; Aucher *et al.*, 2018]. Yet, in this chapter, we will stick mainly with the much simpler pilot systems presented in the preceding sections.

3 Deontic logic as preference logic

Having set up the machinery for changing informational attitudes, we now turn to our major interest in this chapter, the realm of normative evaluation for worlds or actions and the matching dynamic deontic logics. Here we will follow a perhaps not uncontroversial track: our treatment of deontic notions and scenarios will be based on *preference* structure and its changes. We believe that this is a conceptually good way of looking at deontic notions, and at the same time, it lends itself very well to treatment by our earlier methods, since at an abstract level, doxastic plausibility order and deontic betterness order are very similar.[17] The results that follow in the coming sections are largely from [Liu, 2008; Girard, 2008], and [Liu, 2011a].[18]

[17]The stated parallel is also well-known from the deontic literature, for instance, in the works of Hansson or van der Torre cited in our text.

[18]To unclutter notation, henceforth, we will mostly suppress agent indices for modal operators in our languages and their corresponding semantic relations. While we believe that deontic scenarios are very often essentially multi-agent in nature, it is useful to stay with single agent notations as long as these suffice.

Let us say a few more words about the connection between deontic logic and preference, to justify our approach in this chapter. Deontic logic is the logical study of normative concepts such as obligation, prohibition, permission and commitment. This area was initiated by von Wright in [von Wright, 1951] who introduced the logic of absolute obligation. As a reaction to paradoxes with this notion, conditional obligation was then proposed in [von Wright, 1956; von Wright, 1964] and [Fraassen, 1972]. Good reviews systematizing the area are found in [Åqvist, 1987; Åqvist, 1994].

One often thinks of deontic logic as the study of some accessibility relation from the actual world to the set of 'ideal worlds', but the more sophisticated view ([Hansson, 1969; Fraassen, 1973] and [Jackson, 1985]) has models with a binary comparison relation.[19] Such more general comparisons between worlds make sense, for instance, when talking and reasoning about 'the lesser of two evils', or about 'improvement' of some given situation.

This is precisely the ordering semantics we already saw for belief, and it would be tedious to indulge in formal definitions at this stage that the reader can easily construct for herself. Our base view is that of binary *pre-orders* as before, for which we will now use the notation R to signal a change from the earlier plausibility interpretation. As usual, imposing further constraints on the ordering will generate deductively stronger deontic logics. The binary relation R now interprets $O\varphi$ (absolute obligation) as φ *being true in all best worlds*, much like belief with respect to plausibility. Then conditional obligation $O^{\psi}\varphi$ is like conditional belief: φ holds *in the best* ψ*-worlds*.[20]

For further information on deontic logic, we refer to [Åqvist, 1994] and various chapters in this Handbook [Gabbay *et al.*, 2013]. Our emphasis in this chapter will be mainly on interfacing with this field.

As we already noted at the start of this chapter, deontic ordering shows intuitive analogies with the notion of *preference*. One can think of betterness as reflecting the preferences of a moral authority or law-giver, and in the happy Kantian case where agents' duties coincide with their

[19] Hansson has argued early on that von Wright-type deontic logic can be naturally interpreted in terms of a preference relation 'is at least as ideal as' among possible worlds – an ordering that we will call 'betterness' in what follows. This research program in deontic logic is still very much alive today, witness the chapter by Xavier Parent in this Handbook.

[20] There are also more abstract neighborhood versions of this semantics, where the current proposition plays a larger role, in terms of binary deontic betterness relations R^{ψ}, where one can set $\mathfrak{M}, s \models O^{\psi}\varphi$ iff for all t in W with $sR^{\psi}t, \mathfrak{M}, t \models \varphi$.

inclinations, deontic betterness *is* in fact the agent's own preference. We claim no novelty for this line of thought, which was advocated forcefully as early as [Hansson, 1969]. With this twist, we can then avail ourselves of existing studies of preference structure and evaluation dynamics, a line of thinking initiated in [van der Torre, 1997] and [van der Torre and Tan, 1999], though we now take the dynamic-epistemic road.

By way of background to what follows, we note that preference logic is a vigorous subject with its own history. For many new ideas and results in the area, we refer to [Hansson, 2001a] and [Grune-Yanoff and Hansson, 2009], while our final section on related literature is also relevant. What we will do next in this chapter is discuss some major recent developments in the study of preference statics and dynamics, emphasizing those that we see as being of relevance to deontic logic, an area where we will return explicitly later on in this chapter.[21]

4 Static preference logic

In the coming sections, we will discuss basic developments in modal preference logic, starting with its statics, and continuing with dynamics of preference change. Our treatment follows pioneering ideas from [Boutilier, 1994] and [Halpern, 1997], and for the dynamics, we mainly rely on [Benthem *et al.*, 2006] and [Benthem and Liu, 2007].

4.1 General modal preference logic

Our basic models are like in decision theory or game theory: there is a set of alternatives (worlds, outcomes, objects) ordered by a primitive ordering that we dub 'betterness' to distinguish it from richer intuitive notions of preference in common speech.[22]

Definition 14. *A* modal betterness model *is a tuple* $\mathfrak{M} = (W, \preceq, V)$ *with* W *a set of worlds or objects,* $\preceq$ *a reflexive and transitive relation over these, and* V *is a valuation assigning truth values to proposition letters at worlds.*[23]

[21] Preference logic tends to focus on the agents' own preferences, not those of others, but it applies equally well to multi-agent settings such as social choice problems, decisions in games, or moral scenarios, where preference orderings of different agents interact in crucial ways.

[22] To repeat an earlier point, while each agent has her own betterness order, in what follows, merely for technical convenience, we suppress indices wherever we can.

[23] As we said before, we use pre-orders since we want the generality of possibly non-total preferences. Still, total orders, the norm in areas like game theory, provide an interesting specialization for the results in this chapter.

The order relation in these models also induces a strict variant $s \prec t$:

If $s \preceq t$ but not $t \preceq s$, then t is *strictly better* than s.

Here is a simple modal language that can already say a good deal about these structures:

Definition 15. *Take any set of propositional variables Φ, with p ranging over Φ. The* modal betterness language *has this inductive syntax rule:*

$\varphi := \top \mid p \mid \neg\varphi \mid \varphi \wedge \psi \mid \langle\leq\rangle\varphi \mid \langle<\rangle\varphi \mid E\varphi.$

The intended reading of $\langle\leq\rangle\varphi$ is "φ is true in a world that is at least as good as the current world", while $\langle<\rangle\varphi$ says that "φ is true in a world that is strictly better than the current world." In addition, the auxiliary *global existential modality* $E\varphi$ says that "there exists a world where φ is true". Also, as usual, we write $[\leq]\varphi$ for the defined universal modality $\neg\langle\leq\rangle\neg\varphi$, and we use $[<]$ and U for the duals of $\langle<\rangle\varphi$ and E, respectively. Combinations of these modalities can capture a wide variety of binary preference statements comparing propositions, witness the cited literature.

The interpretation of this multi-modal language over our models is entirely standard:

Definition 16. *Truth conditions for the atomic propositions and Boolean combinations are standard. Modalities are interpreted like this:*

- $\mathfrak{M}, s \models \langle\leq\rangle\varphi$ *iff* *for some t wih $s \preceq t$, $\mathfrak{M}, t \models \varphi$.*
- $\mathfrak{M}, s \models \langle<\rangle\varphi$ *iff* *for some t with $s \prec t$, $\mathfrak{M}, t \models \varphi$.*
- $\mathfrak{M}, s \models E\varphi$ *iff* *for some world t in W, $\mathfrak{M}, t \models \varphi$.*

The defined modalities use the obvious universal versions of these clauses. For a concrete glimpse of the reasoning supported by this, we state the standard calculus to come out of this.

Theorem 17. *Modal betterness logic is completely axiomatized by*

1. *the system **S4** for the preference modality,*
2. *the system **S5** for the universal modality,*
3. *the connecting law $U\varphi \rightarrow [\preceq]\varphi$,*
4. *three technical axioms that govern the strict betterness modality and its interaction with the weak preference modality, found in [Benthem* et al.*, 2009c].*

4.2 Special features of preference

Next we briefly survey three special logical features of preference structure that go beyond standard modal logic of pre-orders, and that will eventually turn out to be of interest to deontics as well.

Lifting to generic preferences. While betterness relates specific objects or worlds, preference is often used generically for comparing different *kinds* of things. Ever since [von Wright, 1963], logicians have also studied preferences $P(\varphi, \psi)$ between propositions, viewed as properties of worlds, or of objects.

There is not one such notion, but many, that can be defined by a *lift* of the betterness order among worlds to sets of worlds, cf. [Halpern, 1997; Benthem *et al.*, 2009c; Liu, 2011a]. For instance, compare your next moves in a game, identified with the set of outcomes that they lead to. Which move is 'better' depends on the criterion chosen: maybe we want to go with the one leading to the highest possible outcome, or the one with the highest minimally guaranteed outcome, etcetera.

Such options are reflected in various quantifier combinations for the lifting. In particular, von Wright's classical study emphasized a lifted $\forall\forall$-type preference between sets P, Q:

$$\forall x \in P\ \forall y \in Q\colon\ x \preceq y.$$

A simpler also useful example is the modal $\forall\exists$-type

$$\forall x \in P\ \exists y \in Q\colon\ x \preceq y.$$

This says that for any P-world, there is a Q-world which is at least as good as that ψ-world. In the earlier-mentioned game setting, this stipulation would say that the most preferred moves have the highest maximal outcomes. Unlike the $\forall\forall$-version, this ubiquitous $\forall\exists$ generic preference can be defined in the above modal preference language, using the universal modality ranging over all worlds:

$$P^{\forall\exists}(\varphi, \psi) := U(\psi \to \langle\leq\rangle\varphi).$$

The latter generic preference $P\varphi\psi$, though also just one lift among many, satisfies the usual properties for preference, reflexivity and transitivity: for instance, $P\varphi\psi$ and $P\psi\chi$ imply that $P\varphi\chi$.[24]

[24] Other quantifed stipulations lead to other generic preferences. This proliferation may be a problem (e.g., 'doing what is best' then depends on one's stipulation as to what 'best' means), but there is no consensus in the literature that one can appeal to. A logical approach at least helps make the options clear.

Ceteris paribus clauses. Unlike plausibility, preference ordering seldom comes in pure form: the comparison between alternatives is often entangled with other considerations. Again, games provide an example. Usually, players do not compare moves via the sets of all their possible outcomes, but rather, they compare the *most plausible* outcomes of their moves. This is the so-called *normality sense* of ceteris paribus preference: we do not compare all the φ and ψ-worlds, but only the 'normal ones' in some relevant sense. This belief restriction, observed by many authors, will return in our discussion of doxastic entanglement of preference in Section 8.

But there are also other natural senses of taking a ceteris paribus clause. It was noticed already in [von Wright, 1963] that there is also an 'equality sense' of preference, involving a hidden assumption of *independence*. In that case, one only make comparisons between worlds where some things or issues are held constant, in terms of giving the same truth values to some specified set of atomic propositions, or complex formulas. The logic of equality-based preference is of independent interest, and it has been axiomatized and analyzed in detail in [Benthem *et al.*, 2009c].

Richer preference languages. Modal languages are just one step on a ladder of formalisms for analyzing reasoning practices. It has been claimed that richer languages are needed to faithfully render basic preference notions, cf. [de Jongh and Liu, 2009] on first-order preferences among objects, [Grandi and Endriss, 2009] on first-order languages of social choice, [Benthem *et al.*, 2006] on hybrid modal preference languages for defining backward induction solutions in games, the hybrid modal language of 'desire' and 'freedom' for decision making in [Guo and Sliegman, 2011], or the modal fixed-point languages for games used in [Benthem, 2014]. Though we will mainly use modal formalisms to make the essential points of this chapter, we will mention the relevance of such richer preference formalisms occasionally.

5 World based dynamics of preference change

Now let us look at how given preferences can change. Intuitively, there are many acts and events that can have such an effect. Perhaps the purest form is a radical *command* by some moral authority to do something. This makes the worlds where we act better than those where we do not (cf. [Yamada, 2006], a pioneering study on the dynamics of deontic commands): at least, if we 'take' the order as a legitimate instruction, and change our evaluation accordingly, overriding any pref-

erences that we ourselves might have had. Technically, this dynamics will change a current betterness relation in a model. These phenomena can be studied entirely along the lines already developed in earlier sections for information dynamics.[25]

5.1 Betterness change

[Benthem and Liu, 2007] is a first systematic study of betterness change using methods from dynamic-epistemic logic. The running example in their approach is a weak 'suggestion' $\sharp\varphi$ that a proposition φ be the case. This relatively modest ordering change leaves the set of worlds the same, but it removes any preferences that the agent might have had for $\neg\varphi$-worlds over φ-worlds among these.[26]

Caveat We are not claiming that the technical notion of suggestion as defined here is the most basic action of preference change or deontic change. We start with this system merely as a simple pilot version for our methodology, just as we did with PAL for information change.

The main point to note here is that events with evaluative import can act as triggers that change some current betterness relation on worlds. In particular, a suggestion $\sharp\varphi$ leads to the following model change:

Definition 18. *Given any modal preference model* $(\mathfrak{M}, s)$*, the* suggestion upgrade $(\mathfrak{M}\sharp\varphi, s)$ *has the same domain, valuation, and actual world as* $(\mathfrak{M}, s)$*, but the new preference relations are now*

$$\preceq_i^* = \preceq_i - \{(s,t) \mid \mathfrak{M}, s \models \varphi \text{ and } \mathfrak{M}, t \models \neg\varphi\}$$

In our modal preference models $\mathfrak{M}$, a matching dynamic modality can be interpreted as:

$$(\mathfrak{M}, s) \models [\sharp\varphi]\psi \quad \text{iff} \quad \mathfrak{M}_{\sharp\varphi}, s \models \psi$$

Again, complete dynamic logics exist (cf. [Benthem and Liu, 2007]). The reader will find it particularly useful to scrutinize the key recursion law for preferences after suggestion.[27]

[25] Of earlier treatments, we mention [van der Torre and Tan, 1999], based on the dynamic semantics for natural language proposed in [Veltman, 1996].

[26] Similar operations to suggestion upgrade have come up recently in logical treatments of relevant alternatives theories in epistemology, for the purpose of modeling changes in what agents consider 'relevant' to making or evaluating a knowledge claim. Cf. [Holliday, 2014], [Benthem, 2016a].

[27] Technically, the simplicity of this law reflects the clear analogy between our universal preference modality and the earlier doxastic notion of safe belief.

Theorem 19. *The dynamic preference logic of suggestion is completely axiomatized, over its static base logic, by the following principles:*

1. $\langle\sharp\varphi\rangle p \;\leftrightarrow\; p$
2. $\langle\sharp\varphi\rangle\neg\psi \;\leftrightarrow\; \neg\langle\sharp\varphi\rangle\psi$
3. $\langle\sharp\varphi\rangle(\psi\wedge\chi) \;\leftrightarrow\; (\langle\sharp\varphi\rangle\psi\wedge\langle\sharp\varphi\rangle\chi)$
4. $\langle\sharp\varphi\rangle\langle\leq\rangle\psi \;\leftrightarrow\; (\neg\varphi\wedge\langle\leq\rangle\langle\sharp\varphi\rangle\psi)\vee(\langle\leq\rangle(\varphi\wedge\langle\sharp\varphi\rangle\psi))$
5. $\langle\sharp\varphi\rangle E\psi \;\leftrightarrow\; E\langle\sharp\varphi\rangle\psi$

Similar completeness results are presented in [Liu, 2011a] for dynamic logics that govern many other kinds of normative action, such as the 'strong commands' corresponding to our earlier radical plausibility upgrade. Following the latter instruction, deontically, the agent incorporates the wish of some over-riding authority.

Deontic logicians (or linguists interested in speech acts) will find it easy to come up with many further normative triggers in between weak suggestions and strong commands, but the above-mentioned methods can deal with a wide variety of such proposals.

5.2 Deriving changes in defined preferences

This is an analysis of betterness change and modal statements about it local to specific worlds. But it also applies to the earlier lifted *generic preferences*. As an illustration, consider the $\forall\exists$-lift defined earlier:

Fact 20. *The following equivalence holds for generic $\forall\exists$ preference:*

$$\langle\sharp A\rangle P^{\forall\exists}(\varphi,\psi) \quad \textit{iff} \quad P^{\forall\exists}(\langle\sharp A\rangle\varphi,\langle\sharp A\rangle\psi)\wedge P^{\forall\exists}((\langle\sharp A\rangle\varphi\wedge A),(\langle\sharp A\rangle\psi\wedge A)).$$

We omit the simple calculation for this outcome. Similar results may be obtained for other set liftings such as Von Wright's $\forall\forall$-version.

Finally, the recursive style of dynamic analysis presented here also applies to various forms of ceteris paribus preference, cf. [Girard, 2008].

5.3 General formats for betterness change

Behind our specific examples of betterness change, there lies a much more general theory that works for a wide class of triggering events that change betterness or evaluation order. One widely applicable way of achieving greater generality in this realm uses program expressions from *propositional dynamic logic PDL*.

For instance, suggesting that φ is defined by the program:

$$\sharp\varphi(R) := (?\varphi; R; ?\varphi) \cup (?\neg\varphi; R; ?\neg\varphi) \cup (?\neg\varphi; R; ?\varphi).$$

where R is the given input relation, while the operations $?\varphi$ test whether the relevant proposition φ, or related ones, hold. In particular, the disjunct $(?\varphi; R; ?\varphi)$ means that we keep all old betterness links that run from φ-worlds to φ-worlds.

The preceding definition is equivalent in the dynamic logic PDL to the following more compact program expression

$$\sharp\varphi(R) := (?\neg\varphi; R) \cup (R; ?\varphi).$$

Again this keeps all old R-links as they were, except for deleting those that ran from φ-worlds to $\neg\varphi$-worlds.

Likewise, our plausibility changers for belief revision can be defined in this format. For instance, the earlier 'radical upgrade' is defined by

$$\Uparrow\varphi(R) := (?\varphi; R; ?\varphi) \cup (?\neg\varphi; R; ?\neg\varphi) \cup (?\neg\varphi; \top; ?\varphi)$$

Here the constant symbol $\top$ denotes the universal relation that holds between any two worlds. This program expression reflects the original meaning of the intended transformation: all φ-worlds become better than all $\neg\varphi$-worlds, whether or not they were better before, and within these two zones, the old ordering remains.[28]

Given any PDL program definition of the above sort, one can automatically write recursion laws for the complete dynamic logic of its induced model change, cf. [Benthem and Liu, 2007] for the precise algorithm that computes these axioms. As an illustration, here is the straightforward computation for suggestions:

$$\begin{aligned}
\langle\sharp\varphi\rangle\langle R\rangle\psi \quad &\leftrightarrow \quad \langle(?\neg\varphi; R) \cup (R; ?\varphi)\rangle\langle\sharp\varphi\rangle\psi \\
&\leftrightarrow \quad \langle ?\neg\varphi; R\rangle\langle\sharp\varphi\rangle\psi \vee \langle R; ?\varphi\rangle\langle\sharp\varphi\rangle\psi \\
&\leftrightarrow \quad (\neg\varphi \wedge \langle R\rangle\langle\sharp\varphi\rangle\psi) \vee \langle R\rangle(\varphi \wedge \langle\sharp\varphi\rangle\psi).
\end{aligned}$$

For alternative general formats of ordering change supporting our sort of dynamic logics, we refer to the 'priority update' with event models proposed in [Baltag and Smets, 2008], the general order merge perspective developed in [Benthem, 2006], as well as the still more general 'dynamic dynamic logic' of [Girard *et al.*, 2012].

[28] Conservative upgrades can be dealt with in a similar way. As commands, these leave the agent more of her original preferences: so, differences with radical commands will show up in judgments of 'conditional betterness', as discussed in the literature on conditional obligation: see [Hansson, 1969].

In our view, the practical and theoretical variety of ordering changes for plausibility and preference is not a nuisance, but a feature. It nicely matches the wealth of evaluative actions that we encounter in daily life.

6 Reason-based preferences

Primitive betterness relations among worlds or objects reflect what are called 'intrinsic preferences'. But very often, our preferences have underlying structure, and we compare according to criteria: our preferences are then reason-based, or 'extrinsic'. In this section we develop the latter view, that has motivations in linguistic Optimality Theory, cf. [Prince and Smolensky, 2004], and belief revision based on entrenchment, cf. [Rott, 2003]. This view also occurs in reason-based deontic logic, cf. [Fraassen, 1973], probably the first paper ever to propose the style of thinking in this chapter, [Goble, 2000] and [Jackson, 1985], as we shall see in Section 9 below.

A simple illustration of our approach, that suffices for many natural scenarios, starts with the special case of linear orders for relevant properties that serve as criteria for determining our evaluation of the comparative merits of objects or worlds.

6.1 Priority based preference

The following proposal has many ancestors, among which we mention the treatment in [Friedman and Halpern, 1995],[Rott, 2003]. We follow [de Jongh and Liu, 2009], that starts from a given primitive ordering among propositions ('priorities' among properties of objects or worlds), and then derives a preference among objects themselves.

Definition 21. *A* priority sequence *is a finite linear sequence of formulas written as follows:* $C_1 \gg C_2 \cdots \gg C_n \quad (n \in \mathbb{N})$*, where the* C_m *come from a language describing objects, with one free variable* x *in each* C_m*.*

Definition 22. *Given a priority sequence and objects* x *and* y*, Pref*(x, y) *is defined lexicographically: at the first property* C_i *in the given sequence where* x, y *have a different truth value,* $C_i(x)$ *holds, but* $C_i(y)$ *fails.*

The logic of this framework is analyzed in [de Jongh and Liu, 2009], while a number of concrete applications to deontic logic are developed in [Benthem *et al.*, 2010].

As it happens, this is only one of many ways of deriving a preference ordering from a given priority sequence. A good overview of existing approaches is found in [Coste-Marquis *et al.*, 2004].

6.2 Pre-orders

In general, comparison orders need not be connected, and then the preceding needs a significant generalization. This was done, in a setting of social choice and belief merge, in the seminal paper [Andréka *et al.*, 2002], which we adapt slightly here to the notion of 'priority graphs', based on the treatment in [Girard, 2008], [Liu, 2011b].

The following definitions contain a free parameter for a *language* L that can be interpreted in the earlier modal betterness models $\mathfrak{M}$. For simplicity only, we will take this to be a simple propositional language of properties of worlds, or on another interpretation: objects.

Definition 23. *A* priority graph $\mathscr{G} = \langle P, < \rangle$ *is a strictly partially ordered set of propositions in the relevant language of properties* L.

Here is how one derives a betterness order from a priority graph:

Definition 24. *Let* $\mathscr{G} = \langle P, < \rangle$ *be a priority graph, and* $\mathfrak{M}$ *a model in which the language* L *defines properties of objects. The* induced betterness relation $\preceq_{\mathscr{G}}$ *between objects or worlds is defined as follows:*

$$y \preceq_{\mathscr{G}} x := \forall P{\in}\mathscr{G}((Py \to Px) \vee \exists P'{<}P(P'x \wedge \neg P'y)).$$

Here, in principle, $y \preceq_{\mathscr{G}} x$ requires that x has every property in the graph that y has. But there is a possibility of 'compensation': if y has P while x does not, this is admissible, provided there is some property P' with higher priority in the graph where x does better: x has P' while y lacks it. Clearly, this stipulation subsumes the earlier priority sequences: linear priority graphs lead to lexicographic order.

One can think of priority graphs of propositions in many ways that are relevant to this chapter. In the informational realm, they are hierarchically ordered information sources, structuring the evidence for agents' beliefs. In the normative realm, they can stand for complex hierarchies of laws, or of norm givers with relative authority.

6.3 Static logic and representation theorem

Next, we state an important technical property of this framework, cf. [Friedman and Halpern, 1995], [Liu, 2011b]. It provides two equivalent ways of looking at our basic models.

Theorem 25. *Let* $\mathfrak{M} = (W, \preceq, V)$ *be any modal preference model, without any constraints on its abstract betterness relation. The following two statements are equivalent:*

(*a*) *The relation $y \preceq x$ is a reflexive and transitive order,*

(*b*) *There is a priority graph $\mathscr{G} = (P, <)$ such that, for all worlds $x, y \in W$, $y \preceq x$ iff $y \preceq_{\mathscr{G}} x$.*

This representation theorem says that the general logic of derived extrinsic betterness orderings is still just that of pre-orders. But it can also be seen as telling us that any intrinsic pre-order for preference can be 'rationalized' as an extrinsic reason-based one by adding structure without disturbing the base model as it is.

6.4 Priority dynamics and graph algebra

Now, we have a new locus for more fine-grained preference change: the family of underlying reasons, which brings its own logical structure. For linear priority sequences, relevant changes involve the obvious operations $[^+C]$ of adding a new proposition C to the right, $[C^+]$ of adding C to the left, and various functions $[-]$ dropping first, last or intermediate elements of a priority sequence. [de Jongh and Liu, 2009] give complete dynamic logics for these. Here is one typical valid principe:

$$[^+C]Pref(x, y) \leftrightarrow Pref(x, y) \vee (Eq(x, y) \wedge C(x) \wedge \neg C(y))$$

Operations for changing preferences multiply in the realm of priority graphs, due to their possibly non-linear structure. In this setting an elegant mathematical approach works in terms of a perspicuous algebra of merely two fundamental operations that combine arbitrary graphs:

- $\mathscr{G}_1; \mathscr{G}_2$ adding a graph to another in top position
- $\mathscr{G}_1 \| \mathscr{G}_2$ adding two graphs in parallel.

One can think of this as natural graph operations of 'sequential' and 'parallel' composition. The special case where one of the graphs consists of just one proposition models our earlier simple update actions.

This graph calculus has been axiomatized completely in [Andréka *et al.*, 2002] by algebraic means, while [Girard, 2008] presents a modal-style axiomatization. We display its major principles here, since they express the essential recursion underlying priority graph dynamics.

Here is one case where, as mentioned earlier, a slight language extension is helpful: in what follows, the proposition letter n is a 'nominal' from hybrid logic denoting one single world.

$$\langle \mathscr{G}_1 \| \mathscr{G}_2 \rangle^{\leq} n \;\leftrightarrow\; \langle \mathscr{G}_1 \rangle^{\leq} n \wedge \langle \mathscr{G}_2 \rangle^{\leq} n.$$

$$\langle \mathscr{G}_1 \| \mathscr{G}_2 \rangle^{<} n \;\leftrightarrow\; (\langle \mathscr{G}_1 \rangle^{<} n \wedge \langle \mathscr{G}_2 \rangle^{\leq} n) \vee (\langle \mathscr{G}_1 \rangle^{\leq} n \wedge \langle \mathscr{G}_2 \rangle^{<} n).$$

$$\langle \mathscr{G}_1 ; \mathscr{G}_2 \rangle^{\leq} n \;\leftrightarrow\; (\langle \mathscr{G}_1 \rangle^{\leq} n \wedge \langle \mathscr{G}_2 \rangle^{\leq} n) \vee \langle \mathscr{G}_1 \rangle^{<} n.$$

$$\langle \mathscr{G}_1 ; \mathscr{G}_2 \rangle^{<} n \;\leftrightarrow\; (\langle \mathscr{G}_1 \rangle^{\leq} n \wedge \langle \mathscr{G}_2 \rangle^{<} n) \vee \langle \mathscr{G}_1 \rangle^{<} n.$$

These axioms reduce complex priority relations to simple ones, after which the whole language reduces to the modal logic of weak and strict atomic betterness orders. Hence, this modal graph logic is decidable.

Thus, we have shown how putting reasons underneath agents' preferences (or, for that matter, their beliefs) admits of precise logical treatment, while still supporting the systematic dynamics that we are after.

7 A two-level view of preference

Now we have two ways of looking at preference: one through intrinsic betterness order on modal models, the other through priority structure giving reasons inducing extrinsic betterness orders. One might see this as calling for a reduction from one level to another, but instead, *combining* the two perspectives seems the more attractive option, as providing a richer modeling tool for preference-driven agency.

7.1 Harmony of world order and reasons

In many cases, the two modeling levels are in close harmony, allowing for easy switches from one to the other (cf. [Liu, 2008]):

Definition 26. *Let α: $(\mathscr{G}, A) \to \mathscr{G}'$, with $\mathscr{G}$, $\mathscr{G}'$ priority graphs, and A a new proposition. Let σ be a map from $(\preceq, A)$ to $\preceq'$, where $\preceq$ and $\preceq'$ are betterness relations over worlds. We say that α* induces *σ, if always:*

$$\sigma(\preceq_{\mathscr{G}}, A) \quad = \quad \preceq_{\alpha(\mathscr{G}, A)}$$

Here are two results that elaborate the resulting harmony between two levels for our earlier major betterness transformers:

Fact 27. *Taking a suggestion A is the map induced by the priority graph update $\mathscr{G} \| A$. More precisely, the following diagram commutes:*

$$\begin{array}{ccc} \langle \mathscr{G}, < \rangle & \xrightarrow{\|A} & \langle (\mathscr{G} \| A), < \rangle \\ \downarrow & & \downarrow \\ \langle W, \preceq \rangle & \xrightarrow{\sharp A} & \langle W, \sharp A(\preceq) \rangle \end{array}$$

For a second telling case of harmony in terms of our earlier themes, consider a priority graph $(\mathscr{G}, <)$ with a new proposition A added on top. The logical dynamics at the two levels is now correlated as follows:

Fact 28. *Placing a new proposition A on top of a priority graph $(\mathscr{G}, <)$ induces the radical upgrade operation $\Uparrow A$ on possible worlds ordering models. More precisely, the following diagram commutes:*

$$\begin{array}{ccc} \langle \mathscr{G}, < \rangle & \xrightarrow{A;\mathscr{G}} & \langle (A;\mathscr{G}), < \rangle \\ \downarrow & & \downarrow \\ \langle W, \preceq \rangle & \xrightarrow{\Uparrow A} & \langle W, \Uparrow A(\preceq) \rangle \end{array}$$

Thus the two kinds of preference dynamics, living at different levels of detail in representing scenarios, dovetail well: [Liu, 2011a] has details.

7.2 Correlated dynamics

There are several advantages to working at both levels without assuming automatic reductions. For a start, not all natural operations on graphs have matching betterness transformers at all. An example from [Liu, 2011b] is *deletion* of the topmost elements from a given priority graph. This syntactic operation of removing criteria is not invariant for replacing graph arguments by other graphs inducing the same betterness order, and hence it is a genuine extension of preference change.

But also not all PDL-definable betterness changers from Section 5.3 are graph-definable. In particular, not all PDL transformers preserve the basic order properties of reflexivity and transitivity guaranteed by priority graphs. For a concrete illustration, consider the program

$?A; R$: 'keep the old relation only from where A is true'.

This change does not preserve reflexivity of an order relation R, as the $\neg A$-worlds now have no outgoing relation arrows any more.[29]

All this argues for a policy of co-existence, modeling both intrinsic and extrinsic preference for agents, with reasons for the latter explicitly encoded in priority graphs as an explicit part of the modeling.[30]

[29] $?A; R$ models a refusal to make betterness comparisons at worlds that lack property A. Though idiosyncratic, this seems a bona fide mind change for an agent.

[30] The observations in this section fit well with a general theme in logics of agency today: that of *tracking* dynamic updates operating at finer levels by operations at coarser levels of representing information [Benthem, 2016b]. Tracking is sometimes possible, sometimes it is not, and there are systematic reasons for these phenomena. We will return to this theme in our section on Further Directions.

Coda: Switching perspectives on preference. One might still have a favorite, and think that intrinsic betterness relations merely reflect an agent's raw feelings or prejudices. But the intrinsic-extrinsic contrast is relative, not absolute. If I obey the command of a higher moral authority, I may acquire an extrinsic preference, whose reason is the duty of obeying a superior. But for that higher agent, the same preference may well be intrinsic: "The king's whim is my law". This observation suggests a further theme: namely, transitioning from one perspective to the other. We conclude with a few remarks on realizing this option.

7.3 Additional dynamics: language change

Technically, intrinsic betterness can become extrinsic through a dynamics that has been largely outside the scope of dynamic-epistemic logic so far, that of *language change*. One mechanism here is the proof of the earlier representation result stated in Theorem 25. It partitions the given betterness pre-order into clusters, and if these are viewed as new relevant reasons or criteria, the resulting strict order of clusters is a priority graph inducing the given order. This may look like mere formal rationalization, but in practice, one often observes agents' preferences between objects, and then postulates reasons for them. A relevant source is the notion of 'revealed preference' from the economics literature: cf. [Houser and Kurzban, 2002].

Thus, our richer view of preference also suggests a new kind of dynamics beyond what we have considered so far. In general, reasons for given preferences may have to come from some other, richer language than the one that we started with: and accordingly, we are witnessing a dynamic act of *language creation*.[31]

8 Combining evaluation and information

We have now completed our exposition of information dynamics as well as preference dynamics, which brought its own further topics. What must have become abundantly clear is that there are strong formal similarities in the logic of order and order change in the two realms. We have not even enumerated all of these similarities, but, for instance, all of our earlier ideas and results about reason-based preference also make sense when analyzing evidence-based belief.

[31] For a pioneering study of the importance of language change in the setting of belief revision, cf. [Parikh, 1999].

This compatibility helps with the next natural step we must take. As we said right at the start of this chapter, the major agency systems of information and evaluation do not live in isolation: they interact all the time. A rational agent can process information well in the sense of proof or observation, but is also 'reasonable' in a broader sense of being guided by goals in some intelligible manner.

This *entanglement* of knowledge, belief, and preference is essential to how preference functions,[32] and it shows in many specific settings. We will look at a few cases, and in particular, their impact on the dynamics of preference change.[33] This is where we need a combination of all ideas presented so far: static epistemic, doxastic logic, preference logic and deontic logic, as well as dynamic logics of update actions appropriate to all these notions.

Though we will mainly discuss here how information dynamics influences preference and deontic notions, the opposite influence is equally real. In particular, successful information flow depends on *trust* and *authority*: both clearly deontic notions.[34]

8.1 Generic preference with knowledge

In Section 4.2, we defined one basic generic preference as follows:

$$Pref^{\forall\exists}(\psi, \varphi) := U(\psi \to \langle\leq\rangle\varphi).$$

This refers to possibilities in the whole model, including even those that an agent might know to be excluded. [Benthem and Liu, 2007] defend this scenario in terms of 'regret', but still, there is also a reasonable intuition that preference only runs among situations that are epistemically possible.

This suggests the entangled notion that, for any ψ-world that is *epistemically accessible* to agent a in the model, there is a world which is at least as good where φ is true. This can be written as follows with an epistemic modality:

$$Pref^{\forall\exists}(\varphi, \psi) ::= K_a(\psi \to \langle\leq\rangle\varphi). \qquad (K_{bett})$$

But this entangled notion is not yet what we are after, since we want the 'better world' to be epistemically accessible itself. [Liu, 2009a]

[32] Think of the crucial notion of *expected value* in making decisions which mixes preference and probability as subjective belief.

[33] For a more general discussion of deontic-epistemic entanglement, we refer to [Pacuit *et al.*, 2006], to which we will return later in this chapter.

[34] Following Wittgenstein, [Brandom, 1994] has argued that language use can only be fully understood in terms of commitments that carry rights and obligations.

shows how this cannot be defined in a simple combined language of knowledge and betterness, and that instead, a richer preference formalism is needed with a new *intersection modality* for epistemic accessibility and betterness. The latter entangled notion can be axiomatized, and it also supports a dynamic logic of preference change as before.[35]

8.2 Generic preference with belief

Entanglement becomes even more appealing with generic preference and belief, where the two relational styles of modeling were very similar to begin with. Again, we might start with a mere combination formula

$$Pref^{\forall\exists}(\varphi,\psi) ::= B_a(\psi \to \langle\leq\rangle\varphi). \qquad (B_{bett})$$

This says that, among the most plausible worlds for the agent, for any ψ-world, there exists a world which is at least as good where φ is true.[36]

Again, this seems not quite right in many cases, since we often want the better worlds relevant to preference to stay inside the most plausible part of the model, being 'informational realists' in our desires, not wanting the impossible. To express this, we again need a stronger merge of the two relations by intersection. The key clause for a corresponding new modality then reads like a 'wishful safe belief':

$$\mathfrak{M}, s \models H\varphi \text{ iff } \textit{for all } t \textit{ with both } s \leq t \textit{ and } s \preceq t, \mathfrak{M}, t \models \varphi.$$

The static and dynamic logic of this entangled notion can be determined using the dynamic-epistemic methodology of earlier sections.

8.3 Other entanglements of preference and normality

Entanglements of plausibility and betterness abound in the literature. E.g., [Boutilier, 1994] has models $\mathfrak{M} = (W, \leq_P, \leq_N, V)$ with W a set of possible worlds, V a valuation function and $\leq_P$, $\leq_N$ transitive connected relations $x \leq_P y$ ('y is as good as x) and $x \leq_N y$ ('y is as normal as x). Such models support an operator of *conditional ideal goal* (IG):

[35] An alternative approach would impose *additional modal axioms* that require betterness alternatives to be epistemic alternatives via frame correspondence. However, this style of working puts constraints on our dynamic operations on models that we have not yet investigated systematically.

[36] One might also use a *conditional belief* $B^{\psi}\langle\leq\rangle\varphi$, but to us, the latter logical form seems to express an intuitively less plausible form of entanglement.

$$\mathfrak{M} \models IG^{\psi}\varphi \text{ iff } Max(\leq_P, Max(\leq_N, Mod(\psi))) \subseteq Mod(\varphi)$$

This says that the best of the most normal ψ worlds satisfy φ. Such entangled notions are expressible in the modal logics of this chapter.

Fact 29. $IG^{\psi}\varphi ::= (\psi \wedge \neg\langle B^s\rangle\psi) \wedge \neg\langle <\rangle(\psi \wedge \neg\langle B^s\rangle\psi) \rightarrow \varphi$[37]

Following up on this, now more in the tradition in agency studies in computer science, [Lang *et al.*, 2003] defines the following normality-entangled notion of preference:

Definition 30. $\mathfrak{M} \models Pref^*(\varphi, \psi)$ *iff for all* $w' \in Max(\leq_N, Mod(\psi))$ *there exists* $w \in Max(\leq_N, Mod(\varphi))$ *such that* $w' <_P w$.

This reflects one of the earlier-mentioned 'ceteris paribus' senses of preference, where one compares only the normal worlds of the relevant kinds.[38] Intriguingly, a source of similar ideas on entanglement is the semantics of expressions like "want" and "desire" in natural language, cf. [Stalnaker, 1984; Heim, 1992; Dandelet, 2014].

The preceding notions are similar to our earlier one with an intersection modality, but not quite. They only compare the two most plausible parts for each proposition.

We will give no deeper analysis of all these interesting entangled notions here, but as one small appetizer, we note that we are still within the bounds of this chapter.

Fact 31. *$Pref^*$ is definable in a modal doxastic preference language.*

8.4 Preference change and belief revision

As we have observed already, our treatment of the statics and dynamics of belief and preference shows many similarities. It is an interesting test, then, if the earlier dynamic logic methods for pure cases transfer to belief-entangled notions of preference.

Intuitively, entangled preferences can change because of two kinds of trigger: evaluative acts like suggestions or more general commands, and informative acts changing our beliefs. As an illustration, we quote a result from [Liu, 2008]:

Theorem 32. *The dynamic logic of the above intersective preference H is axiomatizable, with the following essential recursion axioms:*

[37] Here, B^s is an earlier-mentioned modality of *strong belief* that we do not define.

[38] This makes sense, for instance, in the field of epistemic game theory, where 'rationality' means comparing moves by their most plausible consequences according to the player's beliefs and then choosing the best.

1. $\langle \sharp A \rangle \langle H \rangle \varphi \leftrightarrow (A \wedge \langle H \rangle (A \wedge \langle \sharp A \rangle \varphi)) \vee (\neg A \wedge \langle H \rangle \langle \sharp A \rangle \varphi)$.
2. $\langle \Uparrow A \rangle \langle H \rangle \varphi \leftrightarrow (A \wedge \langle H \rangle (A \wedge \langle \Uparrow A \rangle \varphi)) \vee (\neg A \wedge \langle H \rangle (\neg A \wedge \langle \Uparrow A \rangle \varphi)) \vee (\neg A \wedge \langle bett \rangle (A \wedge \langle \Uparrow A \rangle \varphi))$.
3. $\langle A! \rangle \langle H \rangle \varphi \leftrightarrow A \wedge \langle H \rangle \langle A! \rangle \varphi$.

Having intersection modalities for static attitudes may not be all that is needed, though. Importantly, there may also be *entangled triggering events* that do not easily reduce to purely informational or purely evaluative actions, or sequential compositions thereof. Such entangled events, too, can be treated in our style, but we omit details here.[39]

Trade-offs between preference change and information change. Finally, as often in logic, distinctions can get blurred through redefinition. For instance, sometimes, the same scenario may be modeled either in terms of preference change, or as information change. Two concrete examples of such redescription are "Buying a House" in [de Jongh and Liu, 2009] and "Visit by the Queen" in [Lang and van der Torre, 2008]. Important though it is, we leave the study of precise connections between different representations of dynamic entangled scenarios to another occasion.

9 Deontic reasoning, changing norms and obligations

Our analysis of information and preference can itself be viewed as a study of normative discourse and reasoning. However, in this section, we turn to explicit deontic scenarios, and take a look at some major issues concerning obligations and norms from the standpoint of dynamic systems for preference change.[40]

9.1 Triggers for deontic actions and events

Perhaps the most immediate concrete task at hand as a testing ground for our treatment is charting the large variety of deontic notions in daily life. There is still an ongoing debate about identifying what are the major deontic notions and their meanings, witness the recent revival of interest in treating permission as a universal modality on its own in [Anglberger *et al.*, 2015], going back to early proposals in [Benthem, 1979].

[39] For an analogy, see the question scenarios involving conversational triggers that induce parallel information and issue change in [Benthem and Minica, 2009].

[40] Our treatment largely follows [Benthem *et al.*, 2010] and [Benthem *et al.*, 2014].

Likewise, there is a large variety of dynamic deontic actions in daily life that affect normative attitudes and betterness orderings. Frequent normative triggers go far beyond the suggestions and commands that we chose as our examples. For instance, basic deontic acts also include the granting of permissions, or the making of promises and threats – as should be clear from many chapters in this Handbook.[41]

We will not undertake an empirical survey of basic vocabulary here, since this is more of a task for moral philosophers or linguistic experts on normative discourse (see also the beginning of Section 10). Nevertheless, the examples in this chapter should have convinced the reader that a dynamic action perspective on deontic issues is natural, and that much can be done with the tools presented here.

Instead of engaging in further detailed studies of deontic discourse and reasoning, we merely consider a number of general topics and trends that have roots in the deontic literature.

9.2 Unary and dyadic obligation on ordering models

Our static logics heavily relied on binary ordering relations. In fact, deontic logic may have been the first area of philosophical logic to adopt this approach, building on observations from ethics that the deontic notions of obligation, permission and prohibition can be naturally made sense of in terms of an *ideality ordering* $\preceq$ on possible worlds. Here is a quote from [Moore, 1903], found in [Fraassen, 1973], p.6:

> " [...] to assert that a certain line of conduct is [...] absolutely right or obligatory, is obviously to assert that more good or less evil will exist in the world, if it is adopted, than if anything else be done instead."

In this line, the pioneering study [Hansson, 1969] interpreted dyadic obligations 'it is obligatory that φ under condition ψ' on semantic models like ours, using a notion of maximality as in our study of belief:

$$\mathcal{M}, s \models O^{\psi}\varphi \Longleftrightarrow Max(||\psi||_{\mathcal{M}}) \subseteq ||\varphi||_{\mathcal{M}}$$

Depending on the properties of the relation $\preceq$, different deontic logics are obtained here: [Hansson, 1969] starts with a $\preceq$ which is only reflexive, moving then to total pre-orders. This is of course the same idea that has also emerged in conditional logic, belief revision, and the linguistic

[41] For some state of the art work on how to model permissions as first-class citizens in deontic logic, see [Anglberger *et al.*, 2015]

semantics of generic expressions.[42] Variations of this modeling have given rise to various preference-based semantics of deontic logic: see [van der Torre, 1997] for an early useful overview.

Recent developments show the continued vitality of this area. Hansen [2005] is a sophisticated study of conditional obligations in a setting of moral conflicts induced by promises or other deontic actions. Hansen's logic for conditional obligation uses van Fraassen-style reason-based deontic order models while adding ideas reminiscent of 'premise semantics' in the area of conditional logic (as well as later strands in the semantics of non-monotonic logic), and it has a complete axiomatization using non-trivial techniques. Also noteworthy are [Parent, 2014] and [Parent, 2015] which settle several long-standing completeness questions for deontic logics using techniques from non-monotonic logic. This work also clarifies various options for defining 'maximality' and 'optimality' on conditional obligation, and shows how, in a deontic setting, some of the traditional technical fixes (such as the use of the 'limit assumption' in conditional semantics) can be circumvented, or at least, be analyzed in a more satisfactory manner.

In this light, our chapter has taken up an old, but still active, strand in the semantics of deontic reasoning, and then added some recent themes concerning preference: criterion-based priority structure, dynamics of evaluative acts and events, and extended logical languages making these explicit. This seems a natural continuation of deontic logic, while also linking it up with developments in other fields.

9.3 Reasons and dynamics in classical deontic scenarios

The dynamic emphasis in this chapter on changes and their triggering events has thrown fresh light on the study of information and preference-based agency. Deontic logic proves to be no exception to this line of analysis, if we also bring in our treatment of reason-based preference (again we remind the reader of the pioneering [Fraassen, 1973]) – as we shall now demonstrate with a few examples.

The Gentle Murder scenario from [Forrester, 1984], p.194, is a classic of deontic logic that illustrates the basic problem of analyzing 'contrary-to-duty' obligations (CTDs).

[42] A deontic criticism of this account has been that conditional obligation loses antecedent strengthening: [Tan and van der Torre, 1996]. This loss, however, makes sense in our view: non-monotonicity is inherent in the dynamics of information, where the set of most ideal worlds can change under update.

Example 33. *"Let us suppose a legal system which forbids all kinds of murder, but which considers murdering violently to be a worse crime than murdering gently. [...] The system then captures its views about murder by means of a number of rules, including these two:*

1. *It is obligatory under the law that Smith not murder Jones.*
2. *It is obligatory that, if Smith murders Jones, Smith do so gently."*

The priority format of Section 6.1, even just with linear sequences, can represent this scenario in a natural way. Recall that a linear priority sequence $P_1, \ldots, P_n$ combines bipartitions $\{\mathcal{I}(p_i), -\mathcal{I}(p_i)\}$ of the domain of discourse S. Moving towards the right direction of the sequence, ever more atoms p_i are falsified. In a deontic reading, this means that, the more we move towards the right side of the sequence, the more violations hold of morally desirable properties.

Concretely, in the Gentle Murder scenario, the result is two classes of ideality: one class I_1 in which Smith does not murder Jones, i.e., $\mathsf{I}_1 := \neg m$; and another I_2 in which either Smith does not murder Jones or he murders him gently, i.e., $\mathsf{I}_2 := \neg m \vee (m \wedge g)$. The relevant priority sequence $\mathcal{B}$ has $\mathsf{I}_2 \prec \mathsf{I}_1$. Such a sequence orders the worlds via its induced relation $\preceq^{IM}_{\mathcal{B}}$ in three clusters. The most ideal states are those satisfying I_1, worse but not worst states satisfy $\mathsf{V}_1 := \neg \mathsf{I}_1$ but at the same time I_2, and, finally, the worst states satisfy $\mathsf{V}_2 := \neg \mathsf{I}_2$.

With this representation, we can take the scenario one step further.

Example 34. *Consider the priority sequence for Gentle Murder from the preceding Example:* $\mathcal{B} = (\mathsf{I}_1, \mathsf{I}_2)$. *We can naturally* restrict $\mathcal{B}$ *to an occurrence of the first violation by intersecting all formulas in the sequence with* V_1. *Then the first proposition becomes a contradiction, distinguishing no worlds. The best among the still available worlds are those with* $Max^+(\mathcal{B}^{\mathsf{V}_1}) = \mathsf{I}_2 \wedge \mathsf{V}_1$. *A next interesting restriction is* $\mathcal{B}^{\mathsf{V}_2}$, *which describes what the original priority sequence prescribes under the assumption that also the CTD obligation "kill gently" has been violated. In this case we end up in a set of states that are all equally bad.*

This sketch may suffice to show how our approach provides a perspective on the deontic robustness of norms and laws viewed as *CTD* structures: they can still function when transgressions have taken place.[43]

Other major deontic scenarios, such as the Chisholm Paradox, are given reason-based dynamic representations in [Benthem *et al.*, 2014].

[43] Representing *CTD* structures in terms of chains of properties already occurs informally in [Fraassen, 1973]. A formal account is in [Governatori and Rotolo, 2005], with a Gentzen proof calculus manipulating formulae of the type $\varphi_1 @ \ldots @ \varphi_n$ with @

9.4 Typology of change at two levels

We have shown how two-level structure of preference provides a natural medium for modeling deontic notions. Likewise, it yields a rich account of deontic changes. In Section 7, we developed a theory of both informational and evaluative changes, operating either directly on possible world order, or on the priority structure underlying such orders. This two-pronged approach also makes sense here.

As an illustration, we add a temporal twist to the above classical deontic scenario, by 'dynamifying' Gentle Murder.

Example 35. *We start with a priority sequence* $\mathcal{B} = (\neg m)$*. This current deontic state of affairs generates a total pre-order where all* $\neg m$ *states are above all* m *states: "It is obligatory under the law that Smith not murder Jones". Now, we refine this order so as to introduce the sub-ideal obligation to kill gently: "it is obligatory that, if Smith murders Jones, Smith murders Jones gently". In more general terms, we want to model a process of refining legal codes by introducing a contrary-to-duty obligation. Intuitively, this change can happen in one of two ways:*

1. *We refine the given betterness ordering 'on the go' by making a further bipartition of the violation states, putting* $m \wedge g$*-states above* $m \wedge \neg g$*-states. This can be seen as the successful execution of a command of the sort "if you murder, then murder gently".*

2. *We introduce a new law 'from scratch', where* $m \to g$ *is now explicitly formulated as a class of possibly sub-ideal states. This can be seen as the enactment of a new priority sequence* $(\neg m, m \to g)$*.*[44]

In this manner, a CTD sequence can be dynamically created either by uttering a sequence of commands stating what ought to be the case in a sub-ideal situation, or by enacting a new priority sequence.

But in this setting, Theorem 27 from Section 7 applies: in terms of betterness among worlds, the two instructions amount to the same thing! In other words, in well-known scenarios such as this, the same deontic change can be obtained both by refining the order dictated by a given law, and by enacting a new law.

a connective representing a 'sub-ideality' relation. It is an interesting open problem if such a proof-theoretic approach can be related to the more semantically oriented modal logics of this chapter. Incidentally, the same sort of interface questions arise for more recent proof-theoretic approaches to deontic logic, such as the substructural logics for analyzing commands and permissions in [Anglberger *et al.*, 2014].

[44] We have encountered this pattern before in our analysis of priority sequences, since $m \to g$ is equivalent to $\neg m \vee (m \wedge g)$.

Of course, this is just a start. Our discussion of two-level dynamics in Section 7.2 and its possible failures of tracking also suggests new issues. For instance, some well-known changes in laws, such as *abrogation* (a counterpart to the earlier operation of 'graph deletion') have no obvious counterpart at the pure worlds level.

9.5 Norm change

The discussion so far leads up to a more general theme of global dynamics. The problem of *norm change* has recently gained attention from researchers in deontic logic, legal theory, and multi-agent systems.

Approaches to norm change fall into two groups. In syntactic approaches—inspired by legal practice—norm change is an operation performed directly on the explicit provisions in the code of the normative system [Governatori and Rotolo, 2008a; Governatori and Rotolo, 2008b; Boella *et al.*, 2009]. In semantic approaches, however, norm change tends to follow deontic preference order (cf. [Aucher *et al.*, 2009b]). Our initial betterness dynamics on models belonged to the latter group, but our priority methods tie norm change to the former more syntactic level of representation.[45]

More drastic changes of norms and moral codes can be modeled, too, in our framework, using the calculus of priority graphs that we have sketched in Section 6. For an extended discussion of norm change and even legal code change, we refer again to [Benthem *et al.*, 2014].

9.6 Entangled changes

Finally, as observed already in Section 8 on entanglement (cf. [Lang *et al.*, 2003] for a deontic discussion), the dynamic logic connection allows for a unified treatment of two kinds of change that mix harmoniously in deontic scenarios: information change given a fixed normative order, and evaluation change modifying such an order.

Natural deontic scenarios can have deeply intertwined combinations of obligation, knowledge and belief. This point has been acknowledged in the recent literature, and led to combinations of deontic and epistemic modalities, [Aucher *et al.*, 2011; Balbiani and Seban, 2011]. In such a setting, simple operator combinations already express intriguing notions, witness distinctions such as that between KO ('knowing one's duty') versus OK ('having a duty to know'). We add a few more illustrations.

[45] The bridge here is our earlier analysis: obligations defined via ideality and maximality are special kinds of classifications of an Andersonian-Kangerian type.

The first example comes from [Liu, 2011a]. Let us consider the conditional obligation $O^{\psi}\varphi$ again. We can understand the condition ψ as a fact, then $O^{\psi}\varphi$ would express an obligation based on somewhat objective condition. However, fullfilling obligations unavoidably involves agents, hence their epistemic attitude immediately become relevant. With this spirit, we can take the condition at least in the following two sense: (a) an agent *knows* that ψ is true or (b) an agent *believes* that ψ is true. In the former case, we would get much weaker obligation, which in constrast with the stronger obligation obtained in the latter case.

Some sophisticated moral scenarios analyzed in [Pacuit *et al.*, 2006] go even further than simple combination, and point at the further conceptual subtleties arising in a dynamic setting congenial to our main theme in this chapter. These include the distinction between learning new facts that trigger duties, such as accidentally finding out that my neighbor is in distress, or having a duty to know, as happens with the intensive care department of a hospital that is supposed to know the condition of their patients. These issues are interesting and worth pursuing. As far as we know, there has been no sustained systematic analysis yet following up on this work.

Many further deontic themes can be analyzed along the above lines. We refer to [Benthem *et al.*, 2010; Benthem *et al.*, 2014] for a detailed treatment of the Chisholm Paradox, and concrete ways in which priority graph calculus models norm change.

Summary. Taken together, the themes presented in this section show how the logical perspective of this chapter connects with deontic issues, and can throw new light on them. Admittedly, for our style of analysis to work, we do need a few ingredients that are not part of traditional formalizations in deontic logic: in particular, dynamic events, and reasons underlying world ordering. But we believe that such ingredients are not artificial, they are there in the very examples used in the field, provided that we 'mine' their texts for additional dynamic and criterion-based linguistic cues, or just re-analyze the relevant deontic scenarios in these richer terms. The above illustrations may at least have suggested that, and how, this might be done.

10 Further directions

Our main presentation has come to an end. Even so, many relevant roads lead from here. Collecting some points from earlier sections, here are a few active directions where deontic logic meets, or could meet, with current trends in dynamic logics of agency.

10.1 Language, speech acts, and agency

Events that drive information or preference change are often *speech acts* of telling, asking, and so on. Natural language has a sophisticated repertoire of speech acts with a deontic flavour (commanding, promising, allowing, and so on) that invite further logical study, taking earlier studies in meta-ethics and Speech Act Theory (cf. [Searle and Veken, 1985]) to the next level. In particular, such studies will also need a more fine-grained account of the *multi-agency* in dynamic triggers, that has been ignored in this chapter. For instance, things are said by someone to someone, and their uptake depends on relations of authority or trust. Likewise, promises, commands, or permissions are given by someone to someone, and their normative effect depends in subtle ways on who does, and is, what. In particular, [Yamada, 2010] is a pioneering study of this fine-structure of normative action using dynamic-epistemic logic.

10.2 Multi-agency and groups

A conspicuous turn in studies of information dynamics has been a strong emphasis on social scenarios with multi-agent interaction. After all, language use is about communication between different agents, a major paradigm for logic is argumentation between different parties, social behaviour is kept in place by mutual expectations, and so on. In the logics for knowledge, belief, and preference of this chapter, this multi-agent turn can be represented by iteration of single-agent modalities, as in a's knowing that b does, or does not, knows some fact, [Benthem, 2011]. The same is true for games (cf. [Benthem, 2014]), a topic that we will briefly address later.

However, eventually, in a social setting, *groups* must also be taken seriously as new collective actors in their own right. Then we need logics that can deal with notions such as 'common knowledge' or 'distributed knowledge', and their counterparts for beliefs (cf. [Fagin *et al.*, 1995; Meyer and van der Hoek, 1995; Baltag *et al.*, 2019]), or even with more truly collective group-level preferences such as those studied in Social Choice Theory (cf. [Endriss, 2011]).

All these logics of social behavior have dynamic-epistemic extensions in the style of this chapter. State-of-the-art samples can be found in [Seligman *et al.*, 2013; Baltag *et al.*, 2013; Hansen and Hendricks, 2014].

The social turn is highly relevant to deontic logic. From the start, deontic notions and morality seems all about *others*: my duties are usually toward other people, my norms come from outside sources: my boss or a lawgiver.[46] In principle, the methods of this chapter can deal with social multi-agent structure in deontic settings, though much remains to be understood. For instance, it is easy to interpret informational iterations such as $K_a K_b p$, in volving different agents – but what, for instance, is the meaning of an iterated obligation $O_a O_b p$? And beyond this, what would be a group-based 'common obligation': is this more like a propositional common belief, or like a demand for joint action of the group? Other relevant issues in this setting are the entanglement of informational and evaluative acts for groups: cf. [Hartog, 1985; Kooi and Tamminga, 2006; Konkka, 2000], and [Holliday, 2009] on morality as held together by social expectations such as trust. An account of deontically relevant actions for groups will also have to include new operations reminiscent of social choice, such as *belief merge* and *preference merge*, where the priority structures of Section 6 may find new uses, now as a logic-friendly model for social institutions: cf. [Grossi, 2007].

10.3 Games and dependent behavior

Multi-agency involves not just social knowledge, beliefs, and preferences, but also by individual and collective action. All these notions come together concretely in the area of *games*, and hence, not surprisingly, logics of agency have close connections with game theory [Shoham and Leyton-Brown, 2008; Benthem, 2014], being the general mathematical study of strategic behavior and its equilibria.

In the normative realm, actions are as crucial as states of the world – even though actions have been largely ignored in this chapter, for reasons of space. In particular, dependent action is crucial in deontic practice (think of sanctions or rewards), and games are a congenial paradigm. Indeed, many topics in this chapter suggest game-theoretic extensions. In particular, we saw how belief-entangled set lifting is crucial to rational choices made by agents, and this entanglement is typical for games.

[46] The social aspect of deontics has long been explicitly acknowledged by computer scientists working on multi-agent systems: cf. [Meyer, 1988; Wooldridge, 2000], and [Rao and Georgeff, 1991].

Thus, multi-agent versions of our logics have turned out to be a natural tool in the analysis of game solution procedures (cf. [Roy, 2008; Dégremont, 2010; Benthem, 2014]). But preference change also makes sense in games, once we see their preference structure not as a static given, but as something that can evolve dynamically during play. For some first excursions in this direction, see [Liu, 2011a] on the topic of rationalizing preferences in the course of, or after, playing a game.

Another intriguing line worth mentioning are recent uses of deontic logic as a sort of high-zoom level language for our ordinary discourse about 'optimal action', where the precise details of game-theoretic solution procedures such as Backward Induction have been suppressed: cf. [Benthem *et al.*, 2006; Benthem, 2014], and [Roy *et al.*, 2014]. This may well become a major new interpretation for deontic formalisms.

But there is also a converse direction in this contact. Ideas from game theory have started entering deontic logic. One interesting example is the use of standard game solution methods as deliberation procedures for moral judgments in [Loohuis, 2009] and [Tamminga, 2013]. One might even argue that dependent social behavior is the very source of morality, and in that sense, games would be a mandatory next stage after the single-episode driven dynamic logics of this chapter.

10.4 Temporal perspective

Games are one longer-term activity, but deontic agency involves many different processes, some even infinite. The general logical setting here are temporal logics (cf. [Fagin *et al.*, 1995; Parikh and Ramanujam, 2003]) where new phenomena come to the fore. Deontics and morality is not just about single episodes, but about action and interaction over time. Early work in deontic logic already used temporal logics: cf. the pioneering dissertation [Eck, 1981]) where events happen in infinite histories, and obligations come and go. Likewise, in the multi-agent community, logics have been proposed for preferences between complete histories, and planning behavior leading to most desired histories (cf. [Meyden, 1996; Sergot, 2004]). Such temporal logics mesh well with dynamic-epistemic logics (cf. [Benthem *et al.*, 2009a]), with an interesting role for *protocols* as a new object of study, i.e., available procedures for reaching goals. Plans and protocols have a clear normative dimension as well, and thus one would wish to incorporate them into the preference dynamics of this chapter.

Our logics in this chapter described single, or just a few, update or reasoning steps, and the same is true for most scenarios in the deontic

literature. However, the broader horizon of single steps is the temporal process of inquiry in the informational case, and the long-term functioning of society in the normative case. Eventually, studying both aspects together, local and global, seems essential.

10.5 Fine-structure of information

Most dynamic logics for agency, whether about information dynamics or evaluation dynamics, are semantic in nature. The states changed by the process are semantic models. However, in philosophical logic, there has been a continuing debate about the right representation of the *information* used by agents. Semantic information as used in this chapter, though common to many areas, including decision theory and game theory, is coarse-grained, identifying logically equivalent propositions, making agents 'omniscient' at least to that degree – thereby suppressing the activity of logical inference as an information-producing process.

Zooming in on the latter dynamics, agents engage in many activities, such as inference, memory retrieval, introspection, or other forms of 'awareness management' that require a more fine-grained notion of information, closer to syntax. Several dynamic logics of this kind have been proposed in recent years, using ideas from proof theory rather than model theory: cf. [Jago, 2006; Velazquez-Quesada, 2009], and the survey chapter [Benthem and Martínez, 2008] on the different notions of information occurring in modern logic.[47]

But new levels keep appearing. One compromise are the 'evidence models' of [van Benthem & Pacuit, 2011] that generalize the modal logics of this chapter to a 'neighborhood semantics' recording the evidence generating the plausibility ordering on which our modeling of belief was based. While this intermediate level still identifies equivalent propositions in the sense of its weaker base logic, it turns out to support a much richer account of events triggering evidence change: closer, in some ways, to the dynamics of our earlier priority graphs.

Finally, these various levels of representing information are not at odds with each other. Another recent topic is that of 'tracking', cf. [Benthem, 2016b] and [Ciná, 2016], mentioned already in connection with our two-level approach to deontic modeling, where one studies systematically under which conditions updates at a coarser level of representation can faithfully track updates performed at some finer level.

[47] The latter distinguishes even further natural varieties of information that can be found in logic today, such as 'correlation-based' and 'procedural' information.

The same issues of grain level for information make sense in the deontic realm. For instance, our priority graphs with reasons for preferences were syntactic objects than get manipulated by insertions, deletions, permutations, and the like. Significantly, 'reason' is a proof- or argumentation-based term. And indeed, deontic logic has more fine-grained proof-theoretic aspects that would be swept under the rug in a purely semantic approach. As just one illustration, consider the following obvious counterpart to the above-mentioned problem of omniscience. My moral obligations to you cannot reasonably be based on my foreseeing every consequence of my duties or commitments. I owe you careful deliberation, not omniscience.[48] For this and other reasons, there is room for more fine-grained dynamic representations of information and evaluation, closer to deontic syntax – where model theory and proof theory may find interesting ways to meet, for instance, [Tosatto *et al.*, 2012], and [Dong and Gratzl, 2016]

10.6 Digression: numerical strength

While the main theme of this chapter is qualitative approaches, it should be mentioned that there are also numerical approaches to preferences, employing utilities (cf. [Rescher, 1966; Trapp, 1985]) or more abstract 'grades' for worlds (cf. [Spohn, 1988]). Dynamic ideas work in this setting, too, witness the modal logic with graded modalities indicating the strength of preference in [Aucher, 2003], which also defines product update for numerical plausibility models. A stream-lined version in [Liu, 2004] uses propositional constants q_a^m saying that agent a assigns the current world a value of at most m. Our earlier ordering models, both for plausibility and for preference, now get numerical graded versions, with more finely-grained statements of strength of belief and of preference. Dynamic updates can now be defined where we assign values to actions or events, using numerical stipulations in terms of 'product update' from the cited references.[49] More complex numerical evaluation uses *utility* as a fine-structure of preference, and its dynamics can also be dealt with in this style: cf. [Liu, 2004; Liu, 2009b].

While the technical details of these approaches are not relevant here, systems like this do address two issues that seem of great deontic relevance. One is the possibility of comparing not just worlds qua pref-

[48] Likewise, citizens are supposed to know the law, but it would be both unrealistic and unfair to require them to be as well-versed as professional lawyers.

[49] The resulting dynamic logic of numerical evaluation can be axiomatized in the same recursive style as the qualitative systems we discussed in this chapter.

erence, but also *actions*, making sense of the principled distinction in ethics between outcome-oriented and deontological views of obligations and commitments. The other major benefit of a quantitative approach is that we can now study the logic of *how much good* an action does, and accordingly, measure the extent to which we can improve current situations by our actions.

10.7 Probability

Another natural quantitative addition to our analysis would be *probability*. Probabilities measure strengths of beliefs, thereby providing fine-structure to the plausibility orderings that we have worked with. But they can also indicate information that we have about a current process, or a reliability we assign to our observation of a current event.[50] Finally, the numerical factors in probability theory also allow us to mix and weigh various factors in the entangled versions of preference and deontic notions discussed in Section 8. A striking entangled notion is *expected value* in probability theory, whose definition mixes beliefs and evaluation. A unified treatment of logical and probabilistic perspectives in the deontic realm seems a clear desideratum.

11 Appendix: relevant strands in the literature

The themes of this chapter have a long history. For instance, we have pointed at the important connections with belief revision theory and non-monotonic logics throughout. Moreover, while we have followed the dynamic-epistemic approach, there are other proposals in the literature for combining and 'dynamifying' preferences, beliefs, and obligations. In addition to the literature cited already, here are some other relevant lines of work we could not fit into the main line of our presentation.

Computation and agency. [Meyer, 1988] is a pioneering study of deontics from a dynamic viewpoint, reducing deontic logics to suitable dynamic logics. In the same tradition, [Meyden, 1996] takes the deontic logic/dynamic logic interface a step further, studying 'free choice permission' with a new dynamic logic where preferences can hold between actions. Completeness theorems for this enriched semantics then result for several systems. [Pucella and Weissmann, 2004] provide a dynamified logic of permission that builds action policies for agents by adding or deleting transitions. [Demri, 2005] reduces an extension of van der

[50] See [Benthem *et al.*, 2009b] for a rich dynamic epistemic logic of probability.

Meyden's logic to PDL, yielding an EXPTIME decision procedure, and showing how PDL can deal with agents' policies. Preference semantics has also been widely used in AI tasks: e.g., [Wellman and Doyle, 1991] gives a preference-based semantics for goals in decision theory. This provides criteria for verifying the design of goal-based planning strategies, and a new framework for knowledge-level analysis of planning systems. [Horty, 1993] studies commonsense normative reasoning, arguing that techniques of non-monotonic logic provide a better framework than the usual modal treatments. Horty's analysis has a range of applications to conflicting obligations and conditional obligations. Finally, [Lang *et al.*, 2003] propose a logic of desires whose semantics contains two ordering relations of preference and normality, and then interpret "in context A, I desire B" as 'the best among the most normal $A \wedge B$ worlds are preferred to the most normal $A \wedge \neg B$ worlds', providing a new entanglement of preference and normality.

Semantics of natural language. In a line going back to [Spohn, 1988], [Veltman, 1996] presents an update semantics for default rules, locating their meaning in the way in which they modify expectation patterns. This is part of a general program of 'update semantics' for conditionals and other key expressions in natural language. [van der Torre and Tan, 1999] use ideas from update semantics to formalize deontic reasoning about obligations. In their view, the meaning of a normative sentence resides in the changes it brings about in the 'ideality relations' of agents to whom a norm applies. [Zarnic, 2003] uses a simple dynamic update logic to formalize natural language imperatives of the form *FIAT* φ, which can be used in describing the search for solutions of planning problems. [Mastop, 2005] extends the update semantic analysis of imperatives to include third person and past tense imperatives, while also applying it to the notion of free choice permission. [Parent, 2003] outlines a preference-based account of communication, which brings the dynamics of changing obligations for language users to the fore. [Yamada, 2008] distinguishes the illocutionary acts of commanding from the perlocutionary acts that affect preferences of addressees, proposes a new dynamic logic which combines preference upgrade and deontic update, and discusses some deontic dilemmas in this setting.

Philosophical logic. The philosophical study of agency has many themes that are relevant to this paper, often inspired by topics in epistemology or by the philosophy of action. In a direction that is complementary to ours, with belief change as a starting point, [Hansson, 1995] identifies four types of changes in preference, namely revision, contraction,

addition and subtraction, and shows that they satisfy plausible postulates for rational changes. The collection [Grune-Yanoff and Hansson, 2009] brings together the latest approaches on preference change from philosophy, economics and psychology. Following Hansson's work, [Alechina *et al.*, 2013] defines minimal preference change in the spirit of the AGM framework and characterises minimal contraction by a set of postulates. A linear time algorithm is proposed for computing preference changes. In addition, going far beyond what we have discussed in this chapter, Hansson has written a series of seminal papers combining ideas from preference logic and deontic logic, see e.g. [Hansson, 1990b; Hansson, 1990a] and [Hansson, 2001b].

Rational choice theory. Preference is at the heart of decision and rational choice. In recent work at the interface of preference logic, philosophy, and social science, themes from our chapter such as reason-based and belief-entangled preference have come to the fore, with further lines of their own. [Dietrich and List, 2013b] and [Dietrich and List, 2013a] point out that, though existing decision theory gives a good account of how agents make choices given their preferences, issues of where these essential preferences come from and how they can change are rarely studied.[51] The authors propose a model in which agents' preferences are based on 'motivationally salient properties' of alternatives, consistent sets of which can be compared using a 'weighing relation'. Two intuitive axioms are identified in this setting that precisely characterize the property-based preference relations. Starting from similar motivations, [Osherson and Weinstein, 2012a] studies reason-based preference in more complex doxastic settings, drawing on ideas from similarity-based semantics for conditional logic. Essentially, preference results here from agents' comparing two worlds, one having some property and the other lacking it, close to their actual world, and comparing these based on relevant aspects of utility. The framework supports extensive analysis in modal logic, including illuminating results on frame correspondence and axiomatization. [Osherson and Weinstein, 2012b] gives an extension to preference in the presence of quantifiers, while [Osherson and Weinstein, 2014] makes a link between these preference models and deontic logic.

[51] These are of course precisely the two main topics of this chapter: for a more extensive discussion of analogies, cf. also [Liu, 2011a].

12 Conclusion

We have shown how dynamic-epistemic logics can deal with information, knowledge, belief, but also with preference, intrinsic or based on criteria, as well as changes in all these dimensions as events happen and agents act. In doing so, we obtained a suggestive framework for the analysis of deontic notions that links them with many strands in the literature on agency. We also hope to have shown how pursuing this perspective may yield a fresh look at many existing normative scenarios, and may suggest new technical questions about deontic logic as traditionally conceived.

References

[Alechina *et al.*, 2013] N. Alechina, F. Liu, and B. Logan. Minimal preference change. In D. Grossi, O. Roy, and H. Huang, editors, *Proceedings of the 4th International Workshop on Logic, Rationality and Interaction (LORI 2013)*, volume 8196 of *FoLLI-LNCS*, pages 15–26. Springer, 2013.

[Andréka *et al.*, 2002] H. Andréka, M. Ryan, and P-Y. Schobbens. Operators and laws for combining preferential relations. *Journal of Logic and Computation*, 12:12–53, 2002.

[Anglberger *et al.*, 2014] A. Anglberger, H. Dong, and O. Roy. Open reading without free choice. In F. Cariani, D. Grossi, J. Meheus, and X. Parent, editors, *Proceedings of 12th International Conference (DEON 2014)*, 2014, pp.19–32.

[Anglberger *et al.*, 2015] A. Anglberger, N. Gratzl, and O. Roy. Obligation, free choice and the logic of weakest permissions. *The Review of Symbolic Logic*, 8:807–827, 2015.

[Åqvist, 1987] L. Åqvist. *Introduction to Deontic Logic and the Theory of Normative Systems*. Naples: Bibliopolis, 1987.

[Åqvist, 1994] L. Åqvist. Deontic logic. In D. Gabbay and F. Guenthner, editors, *Handbook of Philosophical Logic*, volume 2, pages 605–714. Dordrecht: Kluwer, 1994.

[Aucher *et al.*, 2009a] G. Aucher, P. Balbiani, LF. del Cerro, and A. Herzig. Global and local graph modifiers. *Electronic Notes in Theoretical Computer Science*, 231:293–307, 2009.

[Aucher *et al.*, 2009b] G. Aucher, D. Grossi, A. Herzig, and E. Lorini. Dynamic context logic. In X. He, J. Horty, and E. Pacuit, editors, *Proceedings of the 2nd International Workshop on Logic, Rationality and Interaction (LORI 2009)*, volume 5834 of *FoLLI-LNAI*, pages 15–26. Springer, 2009.

[Aucher *et al.*, 2011] G. Aucher, G. Boella, and L. van der Torre. A dynamic logic for privacy compliance. *Artif. Intell. Law*, 19(2-3):187–231, 2011.

[Aucher *et al.*, 2018] G. Aucher, J. van Benthem, and D. Grossi, Modal logics of sabotage revisited, *Journal of Logic and Computation*, 28(2): 269ÃŘ303, 2018

[Aucher, 2003] G. Aucher. A combined system for update logic and belief revision. Master's thesis, MoL-2003-03. ILLC, University of Amsterdam, 2003.

[Balbiani and Seban, 2011] P. Balbiani and P. Seban. Reasoning about permitted announcements. *Journal of Philosophical Logic*, 40(4):445–472, 2011.

[Baltag and Smets, 2008] A. Baltag and S. Smets. A qualitative theory of dynamic interactive belief revision. In M. Wooldridge G. Bonanno, W. van der Hoek, editor, *Logic and the Foundations of Game and Decision Theory*, volume 3 of *Texts in Logic and Games*. Amsterdam: Amsterdam University Press, 2008.

[Baltag *et al.*, 1998] A. Baltag, L.S. Moss, and S. Solecki. The logic of common knowledge, public announcements, and private suspicions. In I. Gilboa, editor, *Proceedings of the 7th Conference on Theoretical Aspects of Rationality and Knowledge (TARK 98)*, pages 43–56, 1998.

[Baltag *et al.*, 2009] A. Baltag, S. Smets, and J. Zvesper. Keep 'hoping' for rationality: A solution to the backward induction paradox. *Synthese*, 169(2):301–333, 2009.

[Baltag *et al.*, 2011] A. Baltag, N. Gierasimczuk, and S. Smets. Belief revision as a truth tracking process. In K.R. Apt, editor, *Proceedings of Theoretical Aspects of Rationality and Knowledge (TARK 2011)*, 2011. ACM Digital Library, pp.187–190.

[Baltag *et al.*, 2013] A. Baltag, Z. Christoff, J.U. Hansen, and S. Smets. Logical models of informational cascades. In J. van Benthem and F. Liu, editors, *Logic Across the University: Foundations and Application*, pages 405–432. College Publications, London, 2013.

[Baltag *et al.*, 2019] A. Baltag, N. Bezhanishvili, A. Özgün and S. Smets , A Topological Approach to Full Belief, *Journal of Philosophical Logic*, 48: 205–244, 2019

[Benthem and Liu, 2007] J. van Benthem and F. Liu. Dynamic logic of preference upgrade. *Journal of Applied Non-Classical Logic*, 17:157–182, 2007.

[Benthem and Martínez, 2008] J. van Benthem and M. Martínez. The stories of logic and information. In J. van Benthem and P. Adriaans, editors, *Handbook of Philosophy of Information*. Amsterdam: Elsevier, 2008.

[Benthem and Minica, 2009] J. van Benthem and S. Minica. Toward a dynamic logic of questions. In X. He, J. F. Horty, and E. Pacuit, editors, *Proceedings of the 2nd International Workshop on Logic, Rationality and Interaction (LORI 2009)*, volume 5834 of *FoLLI-LNAI*, pages 27–41. Springer, 2009.

[Benthem and Pacuit, 2011] J. van Benthem and E. Pacuit. Dynamic logics of evidence-based beliefs. *Studia Logica*, 99:61–92, 2011.

[Benthem and Smets, 2015] J. van Benthem and S. Smets. Dynamic logics of belief change. In H. van Ditmarsch, J.Y. Halpern, W. van der Hoek, and B. Kooi, editors, *Handbook of Logics for Knowledge and Belief*, pages 313–393. College Publications, 2015.

[Benthem *et al.*, 2006] J. van Benthem, S. van Otterloo, and O. Roy. Preference logic, conditionals and solution concepts in games. In H. Lagerlund, S. Lindström, and R. Sliwinski, editors, *Modality Matters: Twenty-Five Essays in Honour of Krister Segerberg*, pages 61–77. Uppsala Philosophical Studies 53, 2006.

[Benthem *et al.*, 2009a] J. van Benthem, J. Gerbrandy, T. Hoshi, and E. Pacuit. Merging frameworks for interaction. *Journal of Philosophical Logic*, 38(5):491–526, 2009.

[Benthem *et al.*, 2009b] J. van Benthem, J. Gerbrandy, and B. Kooi. Dynamic update with probabilities. *Studia Logica*, 93(1):67–96, 2009.

[Benthem *et al.*, 2009c] J. van Benthem, P. Girard, and O. Roy. Everything else being equal: A modal logic approach for ceteris paribus preferences. *Journal of Philosophical Logic*, 38(1):83–125, 2009.

[Benthem *et al.*, 2010] J. van Benthem, D. Grossi, and F. Liu. Deontics = betterness + priority. In G. Governatori and G. Sartor, editors, *Deontic Logic in Computer Science, 10th International Conference, DEON 2010*, volume 6181 of *LNAI*, pages 50–65. Springer, 2010.

[Benthem *et al.*, 2014] J. van Benthem, D. Grossi, and F. Liu. Priority structures in deontic logic. *Theoria*, 80(2):116–152, 2014.

[Benthem, 1979] J. van Benthem. Minimal deontic logics. *Bulletin of the Section of Logic*, 8:36–41, 1979.

[Benthem, 2006] J. van Benthem. Belief update as social choice. In P. Girard, O. Roy, and M. Marion, editors, *Dynamic Formal Epistemology*, pages 151–160. Springer, Dordrecht, 2006.

[Benthem, 2007] J. van Benthem. Dynamic logic for belief revision. *Journal of Applied Non-Classical Logic*, 17:129–156, 2007.

[Benthem, 2010] J. van Benthem. *Modal Logic for Open Minds*. Stanford: CSLI Publications, 2010.

[Benthem, 2011] J. van Benthem. *Logical Dynamics Information And Interaction*. Cambridge University Press, 2011.

[Benthem, 2014] J. van Benthem. *Logic in Games*. The MIT Press, 2014.

[Benthem, 2016a] J. van Benthem. Talking about knowledge. In L. Moss C. Başkent and R. Ramanujam, editors, *Rohit Parikh on Logic, Language and Society*. Springer, 2016.

[Benthem, 2016b] J. van Benthem. Tracking information. ILLC Research Report PP-2016-02, 2016.

[Blackburn *et al.*, 2001] P. Blackburn, M. de Rijke, and Y. Venema. *Modal Logic*. Cambridge: Cambridge University Press, 2001.

[Blackburn *et al.*, 2007] P. Blackburn, J. van Benthem, and F. Wolter. *Handbook of Modal Logic*. Elsevier, 2007.

[Boella *et al.*, 2009] G. Boella, G. Pigozzi, and L. van der Torre. Normative framework for normative system change. In P.Decker, J.Sichman, C.Sierra, and C.Castelfranchi, editors, *Proceedings of the Eighth International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2009)*, pages 169–176, 2009.

[Boutilier, 1992] C. Boutilier. *Conditional Logics for Default Reasoning and Belief Revision*. PhD thesis, University of Toronto, 1992.

[Boutilier, 1994] C. Boutilier. Conditional logics of normality: A modal approach. *Artificial Intelligence*, 68:87–154, 1994.

[Brandom, 1994] R. Brandom. *Making it Explicit: Reasoning, Representing, and Discursive Commitment*. Harvard University Press, 1994.

[Burgess, 1984] J. Burgess. Basic tense logic. In D. Gabbay and F. Guenthner, editors, *Handbook of Philosophical Logic*, volume 2, pages 89–133. Dordrecht: D. Reidel, 1984.

[Ciná, 2016] G. Ciná, *Categories for the Working Modal Logician*, Dissertation, ILLC, University of Amsterdam.

[Coste-Marquis *et al.*, 2004] S. Coste-Marquis, J. Lang, P. Liberatore, and P. Marquis. Expressive power and succinctness of propositional languages for preference representation. In D. Dubois, C. Welty, and M-A. Williams, editors, *Proceedings of the 9th International Conference on Principles of Knowledge Representation and Reasoning (KR 2004)*. Menlo Park, CA: AAAI Press, 2004. pp.203–212.

[Dong and Gratzl, 2016] H. Dong and N. Gratzl. Open reading for free choice permission: A perspective in substructural logics. Manuscript, 2016.

[Dandelet, 2014] S. Dandelet. Partial desires, blinkered beliefs. Manuscript, Department of Philosophy, University of California at Berkeley.

[de Jongh and Liu, 2009] D. de Jongh and F. Liu. Preference, priorities and belief. In T.Grune-Yanoff and S.O. Hansson, editors, *Preference Change: Approaches from Philosophy, Economics and Psychology*, Theory and Decision Library, pages 85–108. Springer, 2009.

[Dégremont, 2010] C. Dégremont. *The Temporal Mind. Observations on the Logic of Belief Change in Interactive Systems*. PhD thesis, ILLC, University of Amsterdam, 2010.

[Demri, 2005] S. Demri. A reduction from DLP to PDL. *Journal of Logic and Computation*, 15:767–785, 2005.

[Dietrich and List, 2013a] F. Dietrich and C. List. A reason-based theory of rational choice. *Nous*, 47(1):104–134, 2013.

[Dietrich and List, 2013b] F. Dietrich and C. List. Where do preferences come from? *International Journal of Game Theory*, 42(3):613–637, 2013.

[Eck, 1981] J. van Eck. *A System of Temporally Relative Modal and Deontic Predicate Logic and its Philosophical Applications*. PhD thesis, University of Groningen, 1981.

[Endriss, 2011] U. Endriss. Applications of logic in social choice theory. In J. Leite, P. Torroni, Th. Agotnes, G. Boella, and L.van der Torre, editors, *Proceedings of the 12th International Workshop on Computational Logic in Multiagent Systems (CLIMA-2011)*, volume 6814 of *LNAI*, pages 88–91. Springer-Verlag, July 2011. Extended abstract corresponding to an invited talk.

[Fagin *et al.*, 1995] R. Fagin, J.Y. Halpern, Y. Moses, and M.Y. Vardi. *Reasoning about Knowledge*. Cambridge, MA: The MIT Press, 1995.

[Forrester, 1984] J. Forrester. Gentle murder, or the adverbial samaritan. *Journal of Philosophy*, 81:193–197, 1984.

[Fraassen, 1972] B. van Fraassen. The logic of conditional obligation. *Journal of Philosophical Logic*, 1:417–438, 1972.

[Fraassen, 1973] B. van Fraassen. Values and the heart's command. *The Journal of Philosophy*, 70(1):5–19, 1973.

[Friedman and Halpern, 1995] N. Friedman and J. Halpern. Plausibility measures: A user's guide. In P. Besnard and S. Hanks, editors, *Proc. Eleventh Conf. on Uncertainty in Artificial Intelligence (UAI 95)*, Morgan Kaufmann, San Francisco, USA, 1995. pages 175–184.

[Friedman and Halpern, 1997] N. Friedman and Joseph Y. Halpern. Modeling belief in dynamic systems, part I: Foundations. *Artificial Intelligence*, 95(2):257–316, 1997.

[Friedman and Halpern, 1999] N. Friedman and Joseph Y. Halpern. Modeling belief in dynamic systems. part II: Revision and update. *Journal of Artificial Intelligence Research*, 10:117–167, 1999.

[Gabbay *et al.*, 2013] D. Gabbay, J.Horty, X. Parent, R. van der Meyden, and L. van der Torre. *Handbook on Deontic Logic and Normative Systems*, 2013. College Publications: London.

[Gerbrandy, 1999] J. Gerbrandy. *Bisimulation on Planet Kripke*. PhD thesis, ILLC, University of Amsterdam, 1999.

[Girard *et al.*, 2012] P. Girard, J. Seligman, and F. Liu. General dynamic dynamic logic. In Thomas Bolander, Torben Braüner, Silvio Ghilardi, and Lawrence S. Moss, editors, *Advances in Modal Logic*, pages 239–260. College Publications, 2012.

[Girard, 2008] P. Girard. *Modal Logics for Belief and Preference Change*. PhD thesis, Stanford University, 2008.

[Goble, 2000] L. Goble. Multiplex semantics for deontic logic. *Nordic Journal of Philosophical Logic*, 5(2):113–134, 2000.

[Governatori and Rotolo, 2005] G. Governatori and A. Rotolo. Logic of violations: A gentzen system for reasoning with contrary-to-duty obligations. *Australasian Journal of Logic*, 3:193–215, 2005.

[Governatori and Rotolo, 2008a] G. Governatori and A. Rotolo. Changing legal systems: Abrogation and annulment. Part 1: Revision and defeasible theories. In R. van der Meyden and L. van der Torre, editors, *Proceedings of the 9th International Conference on Deontic Logic in Computer Science (DEON 2008)*, volume 5076 of *LNAI*, pages 3–18. Springer, 2008.

[Governatori and Rotolo, 2008b] G. Governatori and A. Rotolo. Changing legal systems: Abrogation and annulment. Part 2: Temporalised defeasible logic. In G. Boella, G. Pigozzi, M. P. Singh, and H. Verhagen, editors, *Proceedings of the 3rd International Workshop on Normative Multiagent System (NorMAS 2008), Luxembourg, Luxembourg, July 14-15, 2008*, pages 112–127, 2008.

[Grandi and Endriss, 2009] U. Grandi and U. Endriss. First-order logic formalisation of Arrow's theorem. In *Proceedings of the 2nd International Workshop on Logic, Rationality and Interaction (LORI-2009)*, volume 5834 of *LNAI*, pages 133–146. Springer-Verlag, October 2009. Also presented at DGL-2009.

[Grossi, 2007] D. Grossi. *Desigining Invisible Hancuffs. Formal Investigations in Institutions and Organizations for Multi-Agent Systems.* PhD thesis, Utrecht University, 2007. SIKS Dissertation Series 2007-16.

[Grove, 1988] A. Grove. Two modelings for theory change. *Journal of Philosophical Logic*, 17:157–170, 1988.

[Grune-Yanoff and Hansson, 2009] T. Grune-Yanoff and S.O. Hansson, editors. *Preference Change: Approaches from Philosophy, Economics and Psychology.* Theory and Decision Library. Springer, 2009.

[Guo and Sliegman, 2011] M. Guo and J. Sliegman. Making choices in social situations. In *Logic and Interactive Rationalityok*, pages 176–202. ILLC, University of Amsterdam, 2011.

[Halpern, 1997] J.Y. Halpern. Defining relative likelihood in partially-ordered preferential structure. *Journal of Artificial Intelligence Research*, 7:1–24, 1997.

[Hansen and Hendricks, 2014] P. G. Hansen and V. F. Hendricks. *Infostorms: How to Take Information Punches and Save Democracy.* Copernicus Books / Springer, 2014.

[Hansen, 2005] J. Hansen. Conflicting imperatives and dyadic deontic logic. *Journal of Applied Logic*, 2:484–511, 2005.

[Hansson, 1969] B. Hansson. An analysis of some deontic logics. *Nous*, 3:373–398, 1969.

[Hansson, 1990a] S.O. Hansson. Defining 'good' and 'bad' in terms of 'better'. *Notre Dame of Journal of Formal Logic*, 31:136–149, 1990.

[Hansson, 1990b] S.O. Hansson. Preference-based deontic logic. *Journal of Philosophical Logic*, 19:75–93, 1990.

[Hansson, 1995] S.O. Hansson. Changes in preference. *Theory and Decision*, 38:1–28, 1995.

[Hansson, 2001a] S.O. Hansson. Preference logic. In D. Gabbay and F. Guenthner, editors, *Handbook of Philosophical Logic*, volume 4, pages 319–393. Dordrecht: Kluwer, 2001.

[Hansson, 2001b] S.O. Hansson. *The Structure of Values and Norms*. Cambridge: Cambridge University Press, 2001.

[Hartog, 1985] G. den Hartog. *Wederkerige Verwachtingen [Mutual Expectations]*. PhD thesis, University of Amsterdam, 1985.

[Heim, 1992] I. Heim. Presupposition projection and the semantics of attitude verbs. *Journal of Semantics*, 9(3):183–221, 1992.

[Holliday, 2009] W.H. Holliday. Dynamic testimonial logic. In X. He, J. Horty, and E. Pacuit, editors, *Proceedings of the 2nd International Workshop on Logic, Rationality, and Interaction (LORI 2009)*, volume 5834 of *FoLLI-LNAI*, pages 161–179. Springer, 2009.

[Holliday, 2012] W.H. Holliday. *Knowing what follows: epistemic closure and epistemic logic*. PhD thesis, Stanford University, 2012.

[Holliday, 2014] W.H. Holliday. Epistemic closure and epistemic logic I: Relevant alternatives and subjunctivism. *Journal of Philosophical Logic*, 2014.

[Horty, 1993] J. Horty. Deontic logic as founded on nonmonotonic logic. *Annals of Mathematics and Artificial Intelligence*, 9:69–91, 1993.

[Houser and Kurzban, 2002] D. Houser and R. Kurzban. Revealed preference, belief, and game theory. *Economics and Philosophy*, 16:99–115, 2002.

[Icard III *et al.*, 2010] T. Icard III, E. Pacuit, and Y. Shoham. Joint revision of beliefs and intentions. In F. Lin, U. Sattler, and M. Truszczynski, editors, *Proceedings of the Twelfth International Conference on the Principles of Knowledge Representation and Reasoning (KR 2010)*, pages 572 – 574. AAAI Publications, 2010.

[Jackson, 1985] F. Jackson. On the semantics and logic of obligation. *Mind*, 94:177–195, 1985.

[Jago, 2006] M. Jago. *Logics for Resource-Bounded Agents*. PhD thesis, University of Nottingham, 2006.

[Konkka, 2000] J. Konkka. Funk games: Approaching collective rationality. E-Thesis, University of Helsinki, 2000.

[Kooi and Tamminga, 2006] B. Kooi and A. Tamminga. Conflicting obligations in multi-agent deontic logic. In J.-J. Ch. Meyer and L. Goble, editors, *Deontic Logic and Artificial Normative Systems: 8th International Workshop on Deontic Logic in Computer Science*, volume 4048 of *LNCS*, pages 175–186. Springer, 2006.

[Lamarre and Shoham, 1994] P. Lamarre and Y. Shoham. Knowledge, certainty, belief, and conditionalisation abbreviated version. In J. Doyle, E. Sandewall, and P. Torasso, editors, *KR'94*, Morgan Kaufmann 1994. pages 415–424.

[Lamarre, 1991] P. Lamarre. S4 as the conditional logic of nonmonotonicity. In J. F. Allen, R. Fikes, and E. Sandewall, editors, *Proceedings of the 2nd International Conference on Principles of Knowledge Representation and Reasoning (KR'91)*, Morgan Kaufmann 1991. pages. 357–367, 1991.

[Lang and van der Torre, 2008] J. Lang and L. van der Torre. From belief change to preference change. In M. Ghallab, C. D. Spyropoulos, N. Fakotakis, and N. Avouris, editors, *Proceedings of the 18th European Conference on Artificial Intelligence (ECAI-2008)*, pages 351–355, 2008.

[Lang *et al.*, 2003] J. Lang, L. van der Torre, and E. Weydert. Hidden uncertainty in the logical representation of desires. In *Proceedings of the 18th International Joint Conference on Artificial Intelligence (IJCAI'03)*, Morgan Kaufmann Publishers Inc. San Francisco, CA, USA. pages 189–231, 2003.

[Lewis, 1973] D. Lewis. *Counterfactuals*. Oxford: Blackwell, 1973.

[Liu *et al.*, 2014] F. Liu, J. Seligman, and P. Girard. Logical dynamics of belief change in the community. *Synthese*, 191:2403–2431, 2014.

[Liu, 2004] F. Liu. Dynamic variations: Update and revision for diverse agents. Master's thesis, MoL-2004-05. ILLC, University of Amsterdam, 2004.

[Liu, 2008] F. Liu. *Changing for the Better: Preference Dynamics and Agent Diversity*. PhD thesis, ILLC, University of Amsterdam, 2008.

[Liu, 2009a] F. Liu. Logics for interaction between preference and belief. Manuscript, Department of Philosophy, Tsinghua University, 2009.

[Liu, 2009b] F Liu. Preference change: A quantitative approach. *Studies in Logic*, 2(3):12–27, 2009.

[Liu, 2011a] F. Liu. *Reasoning about Preference Dynamics*, volume 354 of *Synthese Library*. Springer, 2011.

[Liu, 2011b] F. Liu. A two-level perspective on preference. *Journal of Philosophical Logic*, 40:421–439, 2011.

[Loohuis, 2009] L.O. Loohuis. Obligations in a responsible world. In X. He, J. Horty, and E. Pacuit, editors, *Proceedings of the 2nd International Workshop on Logic, Rationality and Interaction (LORI 2009)*, volume 5834 of *FoLLI-LNAI*, pages 251–262. Springer, 2009.

[Mastop, 2005] R. Mastop. *What Can You Do? Imperative Mood in Semantic Theory*. PhD thesis, ILLC, University of Amsterdam, 2005.

[Meyden, 1996] R. van der Meyden. The dynamic logic of permission. *Journal of Logic and Computation*, 6:465–479, 1996.

[Meyer and van der Hoek, 1995] J.-J.Ch. Meyer and W. van der Hoek. *Epistemic Logic for Computer Science and Artificial Intelligence*. Cambridge: Cambridge University Press, 1995.

[Meyer, 1988] J-J.Ch. Meyer. A different approach to deontic logic: Deontic logic viewed as a variant of dynamic logic. *Notre Dame Journal of Formal Logic*, 29:109–136, 1988.

[Moore, 1903] G. E. Moore. *Principia Ethica*. Cambridge University Press, 1903.

[Osherson and Weinstein, 2012a] D. Osherson and S. Weinstein. Preference Based on Reasons. *The Review of Symbolic Logic*, 5:122–147, 3 2012.

[Osherson and Weinstein, 2012b] D. Osherson and S. Weinstein. Quantified Preference Logic. arXiv:1208.2921, 2012.

[Osherson and Weinstein, 2014] D. Osherson and S. Weinstein. Deontic modality based on preference. arXiv:1409.0824, 2014.

[Otterloo, 2005] S. van Otterloo. *A Strategic Analysis of Multi-agent Protocols*. PhD thesis, Liverpool University, UK, 2005.

[Pacuit *et al.*, 2006] E. Pacuit, R. Parikh, and E. Cogan. The logic of knowledge based on obligation. *Synthese*, 149:311–341, 2006.

[Parent, 2003] X. Parent. Remedial interchange, contrary-to-duty obligation and commutation. *Journal of Applied Non-Classical Logics*, 13(3/4):345–375, 2003.

[Parent, 2014] X. Parent. Maximality vs. optimality in dyadic deontic logic. *Journal of Philosophical Logic*, 43:1101–1128, 2014.

[Parent, 2015] X. Parent. Completeness of Åqvist's systems E and F. *The Review of Symbolic Logic*, 8:164–177, 2015.

[Parikh and Ramanujam, 2003] R. Parikh and R. Ramanujam. A knowledge-based semantics of messages. *Journal of Logic, Language and Information*, 12:453–467, 2003.

[Parikh, 1999] R. Parikh. Beliefs, belief revision, and splitting languages. In J. Ginzburg, L. Moss, and M. de Rijke, editors, *Logic, Language and Computation*, volume 2, pages 266–278. Center for the Study of Language and Information Stanford, CA, 1999.

[Plaza, 1989] J.A. Plaza. Logics of public communications. In M. Emrich, M. Pfeifer, M. Hadzikadic, and Z. Ras, editors, *Proceedings of the 4th International Symposium on Methodologies for Intelligent Systems: Poster Session Program*, pages 201–216, 1989.

[Prince and Smolensky, 2004] A. Prince and P. Smolensky. *Optimality Theory: Constraint Interaction in Generative Grammar*. Oxford: Blackwell, 2004.

[Pucella and Weissmann, 2004] R. Pucella and V. Weissmann. Reasoning about dynamic policies. In I. Walukiewicz, editor, *Proceedings FoSSaCS-7*, Lecture Notes in Computer Science 2987, pages 453–467, 2004.

[Rao and Georgeff, 1991] A.S. Rao and M.P. Georgeff. Modeling rational agents within a BDI-architecture. In J. Allen, R. Fikes, and E. Sandewall, editors, *Proceedings of the 2nd International Conference on Principles of Knowledge Representation and Reasoning*, pages 473–484. San Mateo, CA: Morgan Kaufmann, 1991.

[Rescher, 1966] N. Rescher. Notes on preference, utility, and cost. *Synthese*, 16:332–343, 1966.

[Rott, 2003] H. Rott. Basic entrenchment. *Studia Logica*, 73:257–280, 2003.

[Rott, 2006] H. Rott. Shifting priorities: Simple representations for 27 iterated theory change operators. In H. Langerlund, S. Lindström, and R. Sliwinski, editors, *Modality Matters: Twenty-Five Essays in Honour of Krister Segerberg*, pages 359–384. Uppsala Philosophical Studies 53, 2006.

[Roy *et al.*, 2014] O. Roy, A.J. Anglberger, and N. Gratzl. The logic of best actions from a deontic perspective. In A. Baltag and S. Smets, editors, *Johan van Benthem on Logic and Information Dynamics*, page 657–76. Springer, 2014.

[Roy, 2008] O. Roy. *Thinking before Acting: Intentions, Logic and Rational Choice*. PhD thesis, ILLC, University of Amsterdam, 2008.

[Searle and Veken, 1985] J. R. Searle and D. van der Veken. *Foundations of Illocutionary Logic*. Cambridge: Cambridge University Press, 1985.

[Segerberg, 2001] K. Segerberg. The basic dynamic doxastic logic of AGM. In M.-A. Williams and H. Rott, editors, *Frontiers in Belief Revision*, pages 57–84. Kluwer Academic Publishers, 2001.

[Seligman *et al.*, 2013] J. Seligman, F. Liu, and P. Girard. Facebook and the epistemic logic of friendship. In B. C. Schipper, editor, *Proceedings of the 14th Conference on Theoretical Aspects of Rationality and Knowledge*, pages 229–238, 2013.

[Sergot, 2004] M. Sergot. $(C+)^{++}$: An action language for modelling norms and institutions. Technical Report 8, Department of Computing, Imperial College, London, 2004.

[Shi, 2014] C. Shi. Logics of evidence-based belief and knowledge. Master's thesis, Tsinghua University, 2014.

[Shoham and Leyton-Brown, 2008] Y. Shoham and K. Leyton-Brown. *Multiagent Systems: Algorithmic, Game Theoretic and Logical Foundations*. Cambridge: Cambridge University Press, 2008.

[Shoham, 1988] Y. Shoham. *Reasoning About Change: Time and Causation from the Standpoint of Artificial Intelligence*. Cambridge, MA: The MIT Press, 1988.

[Snyder, 2004] J. Snyder. Product update for agents with bounded memory. Manuscript, Department of Philosophy, Stanford University, 2004.

[Spohn, 1988] W. Spohn. Ordinal conditional functions: A dynamic theory of epistemic states. In W.L. Harper and B. Skyrms, editors, *Causation in Decision, Belief Change and Statistics II*, pages 105–134. Dordrecht: Kluwer, 1988.

[Stalnaker, 1984] R. Stalnaker. *Inquiry*. Cambridge University Press: Cambridge, United Kingdom, 1984.

[Stalnaker, 1996] R. Stalnaker. Knowledge, belief and counterfactual reasoning in games. *Economics and Philosophy*, 12(2):133–163, 1996.

[Stalnaker, 1999] R. Stalnaker. Extensive and strategic forms: Games and models for games. *Research in Economics*, 53:293–319, 1999.

[Stalnaker, 2006] R. Stalnaker. On logics of knowledge and belief. *Philosophical Studies*, 128(1):169–199, 2006.

[Tamminga, 2013] A. Tamminga. Deontic logic for strategic games. *Erkenntnis*, 78(1):183–200, 2013.

[Tan and van der Torre, 1996] Y.-H. Tan and L. van der Torre. How to combine ordering and minimizing in a deontic logic based on preferences. In M. Brown and J. Carmo, editors, *Deontic Logic, Agency and Normative Systems, DEON '96: The 3rd International Workshop on Deontic Logic in Computer Science*, pages 216–232, 1996.

[Tosatto *et al.*, 2012] S. C. Tosatto, G. Boella, L. van der Torre, and S. Villata. Abstract normative systems: Semantics and proof theory. In *Thirteenth International Conference on the Principles of Knowledge Representation and Reasoning*, pages 358–368. AAAI Publications, 2012.

[Trapp, 1985] R.W. Trapp. Utility theory and preference logic. *Erkenntnis*, 22:301–339, 1985.

[van der Torre and Tan, 1999] L. van der Torre and Y.-H. Tan. An update semantics for deontic reasoning. In P. McNamara and H. Prakken, editors, *Norms, Logics and Information Systems*, pages 73–90. Amsterdam: IOS Press, 1999.

[van der Torre, 1997] L. van der Torre. *Reasoning about Obligations: Defeasibility in Preference-based Deontic Logic*. PhD thesis, Rotterdam, 1997.

[van Ditmarsch *et al.*, 2007] H. van Ditmarsch, W. van der Hoek, and B. Kooi. *Dynamic Epistemic Logic*. Berlin: Springer, 2007.

[Velazquez-Quesada, 2009] F. R. Velazquez-Quesada. Inference and update. *Synthese*, 169:283–300, 2009.

[Veltman, 1985] F. Veltman. *Logics for Conditionals*. PhD thesis, University of Amsterdam, 1985.

[Veltman, 1996] F. Veltman. Defaults in update semantics. *Journal of Philosophical Logic*, 25:221–261, 1996.

[von Wright, 1951] G. H. von Wright. Deontic logic. *Mind*, 60:1–15, 1951.

[von Wright, 1956] G. H. von Wright. A note on deontic logic and derived ogligation. *Mind*, 65:507–509, 1956.

[von Wright, 1963] G. H. von Wright. *The Logic of Preference*. Edinburgh: Edinburgh University Press, 1963.

[von Wright, 1964] G. H. von Wright. A new system of deontic logic. In *Danish Yearbook of Philosophy*, pages 173–182. Museum Tusculanum Press, Copenhagen, 1964.

[Wellman and Doyle, 1991] M. Wellman and J. Doyle. Preferential semantics for goals. In *Proceedings of the National Conference on Artificial Intelligence*, pages 698–703, 1991.

[Wooldridge, 2000] M. Wooldridge. *Reasoning about Rational Agents*. Cambridge, MA: The MIT Press, 2000.

[Yamada, 2006] T. Yamada. Acts of commands and changing obligations. In K. Inoue, K. Satoh, and F. Toni, editors, *Proceedings of the 7th Workshop on Computational Logic in Multi-Agent Systems (CLIMA VII)*, 2006. Revised version appeared in LNAI 4371, pages 1-19, Springer-Verlag, 2007.

[Yamada, 2008] T. Yamada. Logical dynamics of some speech acts that affect obligations and preferences. *Synthese*, 165(2):295–315, 2008.

[Yamada, 2010] T. Yamada. Scorekeeping and dynamic logics of speech acts. Manuscript, Hokkaido University, 2010.

[Zarnic, 2003] B. Zarnic. Imperative change and obligation to do. In K. Segerberg and R. Sliwinski, editors, *Logic, Law, Morality: Thirteen Essays in Practical Philosophy in Honour of Lennart Aqvist*, pages 79–95. Uppsala Philosophical Studies 51. Dept. of Philosophy, Uppsala University, 2003.

Johan van Benthem
Institute for Logic, Language and Computation (ILLC),
University of Amsterdam, The Netherlands.
Department of Philosophy, Stanford University, USA.
Department of Philosophy, Tsinghua University, Beijing, China.
Email: j.vanbenthem@uva.nl

Fenrong Liu
Department of Philosophy, Tsinghua University, Beijing, China.
Email: fenrong@tsinghua.edu.cn

More New Frameworks

5
Adaptive Deontic Logics

Frederik Van De Putte, Mathieu Beirlaen and Joke Meheus

Abstract. Adaptive Logics (ALs) are a viable and useful formal tool to handle various aspects of normative reasoning. In this chapter, we motivate, explain, illustrate, and discuss the use of ALs in deontic logic. Published work on deontic ALs focusses mainly on conflicttolerant deontic logics (logics that can accommodate conflicting obligations) and–to a lesser extent–on problems concerning factual and deontic detachment. So does the present chapter. Near the end of the chapter, however, we also indicate some of the possibilities that the adaptive logic framework creates for tackling other types of problems within deontic logic.

1 Introduction **369**

2 Some formal preliminaries **375**

3 Adaptive logics **377**
3.1 The basics 378
3.2 The standard format 385
3.3 Some meta-properties of ALs in standard format 390
3.4 Variants and extensions of the standard format 392
3.5 Further reading 394

4 Revisionist adaptive deontic logics **395**
4.1 **SDL**: three ways of giving it up (while keeping it) 396
4.2 Criteria for comparison and evaluation 398

5 Adaptive inheritance **403**
5.1 Logics with unconflicted inheritance 403
5.2 Evaluating the logics 408
5.3 Further reading and open ends 411

The authors are indebted to Lou Goble, Stef Frijters, Jesse Heyninck, and Daniela Glavaničová for their comments on an earlier version of this chapter.

Handbook of Deontic Logic and Normative Systems, Volume 2

6 Adaptive aggregation **412**
6.1 Adaptive aggregation: a basic example 412
6.2 Evaluating the logics 416
6.3 Further reading and open ends 418

7 Inconsistency-adaptive deontic logics **419**
7.1 Paraconsistent adaptive deontic logic 420
7.2 Semi-paraconsistent adaptive deontic logic 425
7.3 Other paraconsistent negations 427
7.4 Further reading and open ends 429

8 Conflict-tolerant adaptive logics: round-up **431**
8.1 Revisionist deontic adaptive logics: overview 432
8.2 *Prima facie* obligations revisited 434

9 Conditional obligations and adaptive detachment **436**
9.1 Adaptive dyadic deontic logics 437
9.2 Adaptive reasoning with conditionals 441
9.3 Adaptive Characterizations of input/output logic 443

10 Deontic compatibility **447**
10.1 Adaptive logics for deontic compatibility 447
10.2 Further reading 449

11 Summary and outlook **450**

Preludium: Nathan's predicament

One Friday evening, Nathan promises his mother that he will look after his little brother, Ben, on Saturday afternoon so that she can visit her sister. A couple of hours later, Nathan's girlfriend Lisa calls. Being a typical teenager and hopelessly in love, he completely forgets about the promise he made earlier to his mother and agrees with Lisa to go with her to the cinema on Saturday afternoon (to see this cool movie – children under the age of 13 not allowed!) and to go for a veggie burger in the evening. On Saturday, Lisa rings at the door. Almost simultaneously, his mother puts on her coat, meanwhile saying "So, I'll be back by five. Don't forget…". Hearing this, Nathan remembers about *both* promises and immediately realizes what kind of situation he is in. Given his promises, there are several things he ought to do and it is clear that he cannot do them all. Keeping his promise to go for a veggie burger in the evening still seems feasible, but he cannot look

after six year old Ben and at the same time take Lisa to this particular movie!

1 Introduction

Logical principles may fail to apply under certain conditions, and logical principles involving normative concepts are no exception. Even if we restrict our focus to the modalities "it is obligatory that" and "it is permitted that", there are circumstances in which we cannot apply certain plausible rules of inference (unrestrictedly) on pain of highly undesirable outcomes or even plain triviality.

The example from the *preludium* provides one kind of illustration of this phenomenon. It concerns a context in which an agent, in this case Nathan, faces several obligations that cannot be jointly fulfilled. In such contexts, several clusters of otherwise plausible principles involving obligations and permissions are problematic. Let us look at two instances of such clusters.

Consider first the combination of the principle that whatever is obligatory is also permissible (OIP), and the principle of the interdefinability of obligation and permission (ID):

(OIP) If A is obligatory, then A is also permitted: $\mathsf{O}A \supset \mathsf{P}A$
(ID) A is obligatory iff $\neg A$ is not permitted: $\mathsf{O}A \equiv \neg\mathsf{P}\neg A$

If both A and its negation $\neg A$ are obligatory ($\mathsf{O}A \wedge \mathsf{O}\neg A$), then by (ID) and the first conjunct, $\neg\mathsf{P}\neg A$. However, by (OIP) and the second conjunct, $\mathsf{P}\neg A$. So we obtain a plain contradiction: $\neg\mathsf{P}\neg A \wedge \mathsf{P}\neg A$. Even if one is willing to accept that contradictions are not absurd, it seems hard to accept that conflicting obligations entail them. Opinions may differ on which of these two principles is the most salient one. It is clear, however, that at least one of them has to be abandoned or adequately restricted if we want to avoid the outcome that conflicting obligations entail plain contradictions.

A second cluster of principles which is problematic in the face of conflicting obligations consists of the aggregation principle (Agg), the principle that "ought implies can" (OIC), and the impossibility of contradictory states of affairs (CP):

(Agg) If A and B are obligatory, then so is their conjunction: $(\mathsf{O}A \wedge \mathsf{O}B) \supset \mathsf{O}(A \wedge B)$
(OIC) If something is obligatory, then it is also possible: $\mathsf{O}A \supset \Diamond A$

(CP) Contradictions are impossible: $\neg\Diamond(A \wedge \neg A)$

If $\mathsf{O}A \wedge \mathsf{O}\neg A$, then, by (Agg), $\mathsf{O}(A \wedge \neg A)$ and hence by (OIC), $\Diamond(A \wedge \neg A)$. But this is in direct contradiction with (CP). Again, one of the principles from the cluster cannot be upheld (unrestrictedly) if we are to accommodate conflicting obligations, or at least if we want to avoid that such conflicts result in plain contradictions.

Besides conflicting obligations, there are other types of circumstances in which plausible logical principles may fail to apply. One that we want to consider here concerns the violation of conditional obligations, i.e. statements of the form "If A is the case, then B is obligatory" – formally, $\mathsf{O}(B \mid A)$. Each of the rules of factual detachment (FD) and deontic detachment (DD) is intuitively appealing as a rule for detaching unconditional obligations from conditional ones:

(FD) If it is obligatory that B given condition A, and if A is the case, then it is obligatory that B: $A, \mathsf{O}(B \mid A) \vdash \mathsf{O}B$

(DD) If it is obligatory that B given condition A, and if A is obligatory, then it is obligatory that B: $\mathsf{O}A, \mathsf{O}(B \mid A) \vdash \mathsf{O}B$.

The combination of (FD) and (DD) is known to cause trouble in so-called contrary-to-duty cases: cases in which a secondary obligation kicks in once a possibly conflicting primary obligation was violated. The following is an example of such a case.

Lisa and Nathan are a couple since eleven months. Lisa wants their first anniversary to be special and promises Nathan to take him to a "real" restaurant. One can only pay in cash at this restaurant, so if they are going to the restaurant, then Lisa ought to withdraw one hundred dollars at an ATM beforehand. However, on the day of the event, Lisa changes her mind and decides that she is not going to the restaurant after all – perhaps she is no longer sure she wants to be Nathan's girlfriend in the first place. In view of her promise, she (still) has the obligation to take Nathan to the restaurant: $\mathsf{O}A$. She also still has the conditional obligation that, if she takes Nathan there, she has to withdraw the money: $\mathsf{O}(B \mid A)$. However, if she is not going to any restaurant, then she should not withdraw a hundred dollars, since carrying around that much money for no reason would be hazardous: $\mathsf{O}(\neg B \mid \neg A)$. And as it happens to be, she is not going to the restaurant: $\neg A$.

Let us now see how the combination of (FD) and (DD) causes trouble for cases like this. If the obligation $\mathsf{O}A$ is violated, i.e. $\neg A$ is the case, then the primary conditional obligation $\mathsf{O}(B \mid A)$ leads to the unconditional obligation $\mathsf{O}B$ via (DD), while the secondary (contrary-to-duty)

obligation $\mathsf{O}(\neg B \mid \neg A)$ leads to the unconditional obligation $\mathsf{O}\neg B$ via (FD). In order to resolve this conflict, we must block the application of (DD) or that of (FD).[1]

We will have much more to say about conflicting obligations and about the detachment of conditional obligations in the remainder of this chapter. For now, these examples merely serve to illustrate a general point. In the circumstances described above – conflicting obligations and contrary-to-duty cases – one cannot consider principles such as the ones just mentioned as unrestrictedly valid. This leaves the logician who wants to explicate our reasoning in such cases with various options. One is to simply reject those principles, and hence declare a number of intuitive inferences simply invalid. Our stance towards this option is perhaps best summarized by the following words of van Benthem [2004, p. 95]:

> This is like turning down the volume on your radio so as not to hear the bad news. You will not hear much good news either.

A more promising option is to look for restricted versions or alternative, more fine-grained formulations of those principles. For instance, for the case of conflicting obligations, one may argue that (Agg) should only be applicable in case the conjunction of A and B is possible. For contrary-to-duty cases, one may reformulate (FD) as a principle that concerns dynamic updates, rather than (mere) factual input – see e.g. [van Benthem *et al.*, 2014] where this is proposed.

We will not pursue this second option here, even though occasionally we will show that some concrete instances of it fail to deliver an appropriate logic of normative reasoning, either on philosophical or on purely technical grounds. Instead, we will focus on a third option, i.e. to take (some of) these problematic principles to be only valid in a defeasible, context-sensitive way.

That this option seems well in line with our intuitions is easily demonstrated by returning to our examples. As soon as Nathan realizes that looking after Ben is incompatible with going to that particular movie with Lisa, it seems quite rational to reject the conclusion that he ought to do both. But, suppose that his mother also made him promise to walk the family dog on Saturday evening. Would it be rational that,

[1] Alternatively, we could bite the bullet and accept the outcome that both B and $\neg B$ are obligatory. But then our first illustration shows that we must give up other logical principles on pain of contradiction.

in view of the conflict concerning his afternoon plans, he also rejects the conclusion that he ought to go with Lisa for a veggie burger (at 6pm) and take the dog for a walk (at 10pm)? It seems that the one should have no bearing on the other. What this comes to is that, even if it makes sense to withdraw applications of (Agg) upon realizing that A and B are mutually exclusive, this need not affect other applications of (Agg).

In a similar vein, it seems quite natural that certain applications of (DD) are upheld *unless and until* it turns out that the unconditional obligation in the premises is violated. That Lisa has the obligation to withdraw money, even if she is not going to the restaurant at all, feels contra-intuitive to non-logicians. Is there something wrong with their intuitions? Not necessarily, and maybe even to the contrary. It seems quite justified that in cases like this, (DD) is treated as a *defeasible* rule of inference: the obligation is detached from the conditional obligation *provided* the unconditional obligation is not violated.

Note the difference between the third option and the first one. In our approach, we do not invalidate *principles*, we invalidate certain *applications* of principles and this is done only *when and where* necessary. This at once illustrates what we mean by context-sensitivity: whether an application of a certain principle or rule is validated or not depends on the specific context (the premises at issue).

The aforementioned clusters of principles governing obligations and permissions were originally introduced to hold unconditionally. The circumstances in which these principles are not (jointly) applicable, such as conflicts and violations, are often considered anomalous or exceptional. Other principles were acknowledged to be applicable only in a defeasible, context-sensitive manner right from their very introduction. We give only one example. Consider the *nullum crimen sine lege* principle: "If A is not forbidden, then A is permitted". This principle is best thought of as a kind of *default* rule: assume (or infer) $\mathsf{P}A$, unless $\mathsf{O}\neg A$ follows from the premises. This rule is defeasible by its very nature, in the sense that at least some of its instances are violated in every interesting application context.

In order to apply inference rules in a logic in a context-sensitive, defeasible manner, the consequence relation of this logic has to be *non-monotonic*: given a set of premises from which a conclusion A is derivable, it must be possible to revoke A in the light of additional premises.[2]

[2]Formally, a logic **L** is non-monotonic iff (if and only if) there are two sets of formulas Γ and Δ and there is a formula A such that A is **L**-derivable from Γ, while A is *not* **L**-derivable from $\Gamma \cup \Delta$.

Adaptive logics (henceforth, ALs) provide a natural way to explicate the premise-sensitive, defeasible application of certain inference rules in a formal logic.

ALs are built on top of a core logic, called the *lower limit logic*, the inference rules of which hold unconditionally and unrestrictedly. An AL strengthens its lower limit logic by allowing a number of additional inference rules to be applied relative to the specific premises at hand. The term "adaptive logic" originates from this premise-sensitivity: ALs "adapt" themselves to the premises under consideration.

Beside ALs, many other formalisms for modelling defeasible reasoning have been applied in a deontic context: default logic [Horty, 2012], defeasible deontic logic [Nute, 1999], Governatori *et al.*'s chapter 9 of this volume, formal argumentation theories [Gabbay, 2012; Prakken and Sartor, 2015; Straßer and Arieli, 2019; Beirlaen and Straßer, 2016; van der Torre and Villata, 2014], input/output logic [Parent and van der Torre, 2013], etc. These different frameworks are all linked to one another and to ALs in various ways – see e.g. [Heyninck and Straßer, 2016] for some recent comparisons.

There is, however, a distinctive feature of ALs that sets them apart from other approaches to non-monotonic reasoning, viz. their dynamic proof theory. The idea behind this proof theory is that the non-monotonicity of the logic's consequence relation is pushed into the object-level proofs. This means that a given derivation in a proof can become rejected in the light of other derivations within that same proof.[3]

Another important difference between the existing work on ALs and other types of non-monotonic logics is the pivotal role that classical logic (henceforth **CL**) plays within the latter. ALs are, at least in origin, more pluralistic in spirit regarding the meaning of the classical connectives, thus opening up to new perspectives on defeasible reasoning that are hard to detect when one sticks to **CL** as one's underlying monotonic logic.[4]

The current chapter's aim is to motivate, explain, illustrate, and discuss the use of ALs in deontic logic. Published work on deontic ALs focusses mainly on conflict-tolerant deontic logics (logics that can accommodate conflicting obligations) and – to a lesser extent – on problems concerning factual and deontic detachment. So does the present chapter. Near the end of the chapter, however, we also indicate some of the possibilities that the adaptive logic framework creates for tackling other

[3]We will define and illustrate the dynamic proof theory of ALs in Section 3.

[4]This aspect of ALs is nicely illustrated by our Section 7, where we introduce and discuss (adaptive) paraconsistent deontic logics.

types of problems within deontic logic.

The outline of this chapter is as follows. For ease of reference, we start by recalling the basic definitions concerning Standard Deontic logic, henceforth **SDL** (Section 2). In Section 3 we provide an introduction to the framework of ALs. By way of illustration, we first present two very simple adaptive logics that can handle examples as the one from the preludium (Section 3.1).

In Sections 5–7 we present and discuss a variety of conflict-tolerant deontic ALs that move further away from the standard view: unlike the logics from Section 3.1, the logics from Sections 5–7 have lower limit logics that are inferentially weaker than **SDL**. Section 4 provides the conceptual and technical basis for this discussion. Whereas Sections 5 and 6 are mainly based on existing work, Section 7 presents mostly new material that we think improves on the existing work in a number of ways – we explain this in Section 7.4.

Section 8 summarizes the merits and demerits of the conflict-tolerant ALs presented throughout Sections 3–7. In that section we also show how the simple logics introduced in Section 3.1 can be further refined in various ways.

The other main application of existing deontic ALs concerns the problem of detaching conditional obligations. We distinguish between various approaches to this problem in Section 9, and discuss adaptive versions of each of them.

In Section 10 we show how the *nullum crimen sine lege* principle can be captured within the AL framework, and how this gives rise to various extensions of the logics defined in previous sections. This at once paves the way for our last section in which we give a short summary of the chapter and point to ideas for future research.

Throughout this chapter our focus is on the illustration and motivation of the core ideas we present, rather than on formal details and meta-theoretical results. Whenever relevant, we provide pointers to the literature, cf. the subsections "further reading and open ends".

Much of what we will write in this chapter builds on Lou Goble's work on normative conflicts, which is nicely summarized in his contribution to the first volume of this handbook, [Goble, 2013]. We will provide references to specific parts of this (and other) work in due course. In general, we try to avoid overlap as much as possible, but whenever this maxim conflicts with keeping the present chapter self-contained, we give priority to the latter.

We end this section with some more general comments regarding the plurality and diversity of logics to be discussed in this chapter. Our

stance on the matter can be described as follows.

For a start, various logics present themselves as useful depending on the specific type of application context, and the associated logical grammar one wants to study. But even if we keep the grammar fixed, there are various reasons for occupying oneself with not one but many logics for this grammar. That logic – even the logic of our most basic connectives like conjunction – is not god-given, and that there are no absolute grounds for preferring one logic over another, seems hardly contested nowadays. So all one can do is give pragmatic arguments, referring to general desiderata for logics on the one hand, and the needs of a given application on the other.

In the context of conflict-tolerant deontic logics, one way to argue for diversity is by referring to various explosion principles, as discussed in Section 4.2. For instance, if one does not *need* to accommodate conflicts between obligations and permissions, or if one can safely assume within a given domain that norms are at least internally consistent, then this should translate to one's preferred logic for that domain. Moreover, there are many different ways one can interpret the O of a given (conflict-tolerant or other) deontic logic, which will yield different formal semantics and hence different logics in turn.

Going non-monotonic (or in our case, going adaptive) does not reduce this plurality – quite to the contrary. To use Makinson's words [2005, p. 14]:

> Leaving technical details aside, the essential message is as follows. Don't expect to find *the* nonmonotonic consequence relation that will always, in all contexts, be the right one to use. Rather, expect to find several *families* of such relations, interesting *syntactic conditions* that they sometimes satisfy but sometimes fail, and principal *ways of generating* them mathematically from underlying structures.

Indeed, it will become clear throughout this chapter that there are usually several interesting and sensible ways of going adaptive, starting from a given lower limit logic. In the absence of further philosophical arguments against the resulting logics, one needs to keep an open mind and study all of them.

2 Some formal preliminaries

Languages Throughout this chapter, we use $A, B, \ldots$ as metavariables for formulas of a given formal language, and $\Gamma, \Delta, \ldots$ as metavariables

for sets of such formulas.

Let henceforth **CL** stand for the propositional fragment of classical logic, as based on a set of propositional variables (also called sentential letters) $\mathcal{S} = \{p, q, \ldots\}$, the connectives $\neg, \vee, \wedge, \supset, \equiv$, and the logical constants $\bot, \top$. We use $\mathcal{W}$ to denote the set of well-formed formulas in this language.

The language of **SDL** is obtained by adding to the grammar of **CL** the modal operators O for "it is obligatory that" and P for "it is permitted that". We take both O and P (and the classical connectives) to be primitive by default in this chapter; i.e. whenever one is defined in terms of the others in one logic or another, we will indicate so. For the sake of simplicity, we will focus on the fragment of this language in which no nested occurrences of O and P are allowed. This means that the set of well-formed formulas for **SDL** is defined as follows:

$$\begin{aligned}\mathcal{W}^d := \quad & \mathcal{W} \mid \neg\langle\mathcal{W}^d\rangle \mid \langle\mathcal{W}^d\rangle \vee \langle\mathcal{W}^d\rangle \mid \langle\mathcal{W}^d\rangle \wedge \langle\mathcal{W}^d\rangle \mid \langle\mathcal{W}^d\rangle \supset \langle\mathcal{W}^d\rangle \mid \\ & \langle\mathcal{W}^d\rangle \equiv \langle\mathcal{W}^d\rangle \mid \mathsf{O}\langle\mathcal{W}\rangle \mid \mathsf{P}\langle\mathcal{W}\rangle\end{aligned}$$

Axiomatization The logic **SDL** is obtained by adding to **CL** the following axioms, rule, and definition:

(K) $\mathsf{O}(A \supset B) \supset (\mathsf{O}A \supset \mathsf{O}B)$
(D) $\mathsf{O}A \supset \neg\mathsf{O}\neg A$
(N) if $\vdash A$, then $\vdash \mathsf{O}A$
(Def$_\mathsf{P}$) $\mathsf{P}A =_{\mathsf{df}} \neg\mathsf{O}\neg A$

It is well-known that in the presence of (N), (K) can equivalently be expressed as the combination of the axiom of aggregation (Agg) and the rule of inheritance (Inh):

(Agg) $(\mathsf{O}A \wedge \mathsf{O}B) \supset \mathsf{O}(A \wedge B)$
(Inh) if $\vdash A \supset B$, then $\vdash \mathsf{O}A \supset \mathsf{O}B$

whence **SDL** can be equivalently characterized by adding (N), (Agg), (Inh), (D), and (Def$_\mathsf{P}$) to **CL**. Note also that in the presence of (Agg), (D) is equivalent to the following principle:

(P) $\neg\mathsf{O}(A \wedge \neg A)$

For ease of reference, we note some more derivable principles of **SDL**. The first is the axiom of distributivity (of O over $\wedge$):

(Dist) $\mathsf{O}(A \wedge B) \supset (\mathsf{O}A \wedge \mathsf{O}B)$

Second, the replacement of equivalents rule (RE) is an immediate consequence of the behavior of $\supset$ and $\equiv$ in **CL** and (Inh):

(RE) if $\vdash A \equiv B$, then $\vdash \mathsf{O}A \equiv \mathsf{O}B$

Third and last, in view of (Agg), (Inh), and the validity of disjunctive syllogism (DS) in **CL**, we have:

(DDS) $(\mathsf{O}A \wedge \mathsf{O}(\neg A \vee B)) \supset \mathsf{O}B$

Semantics We work with the traditional Kripke-semantics for **SDL**, but to prepare for the semantics of other logics to be presented below, we work with a designated "actual" world. An **SDL**-model M is a quadruple $\langle W, w_0, R, v\rangle$, where W is a non-empty set of worlds, $w_0 \in W$ is the actual world, $R \subseteq W \times W$ is a serial[5] accessibility relation and $v : W \to \mathcal{S}$ is a valuation function. $R(w)$ (the image of w under R) is the set of worlds that are accessible from the viewpoint of w, $R(w) = \{w' \mid (w, w') \in R\}$.

The semantic clauses for the sentential variables and the connectives are as usual; those for O and P are as follows:

(SC1) $M, w \models \mathsf{O}A$ iff $M, w' \models A$ for all $w' \in R(w)$
(SC2) $M, w \models \mathsf{P}A$ iff $M, w' \models A$ for some $w' \in R(w)$

Truth of a formula A at a world w is given by the relation $\models$. Truth in a model $M = \langle W, w_0, R, v\rangle$ is simply truth at w_0. We say M is a model of Γ iff all the members of Γ are true in M, i.e. if for all $B \in \Gamma$, $M, w_0 \models B$. Semantic consequence is then defined as the preservation of truth in all models: $\Gamma \Vdash A$ iff A is true in all models of Γ.

Following customary notation, let $|A|_M =_{\mathsf{df}} \{w \mid M, w \models A\}$. $|A|_M$ is also called the *truth set* (intension) of A. Note that the semantic clause for O can be equivalently rewritten as follows: $M, w \models \mathsf{O}A$ iff $R(w) \subseteq |A|_M$.

3 Adaptive logics

Adaptive logics were originally introduced by Diderik Batens around the 1980s, and have since been applied to various forms of defeasible reasoning.[6] The aim of this section is to highlight the basic features of ALs by means of a running example, viz. the logics $\mathbf{SDL}^{\mathbf{r}}_{\mathbf{p}}$ and $\mathbf{SDL}^{\mathbf{m}}_{\mathbf{p}}$.

[5] R is serial iff for every $w \in W$, there is a $w' \in W$ such that $(w, w') \in R$.
[6] See Section 3.5 for references to the literature on ALs.

These logics can handle simple cases of conflicting obligations such as the running example from the beginning of this chapter. We explain the idea behind both logics in Section 3.1. Generic definitions for all ALs in the standard format from [Batens, 2007] are given in Section 3.2. We mention the most salient properties of all logics that are defined within this format in Section 3.3. Finally, we discuss some variants of the standard format that will turn out useful in the remainder of this chapter (Section 3.4).

3.1 The basics

Before introducing the logics $\mathbf{SDL^r_p}$ and $\mathbf{SDL^m_p}$, we present another predicament from Nathan's life. The example will be used to illustrate the proof theory of $\mathbf{SDL^r_p}$ and $\mathbf{SDL^m_p}$.

One evening, Nathan comes home from school. As soon as he enters the kitchen, he hears his father: "Remember, Nathan, it's your turn to do the dishes tonight. Do them this time!" His mother immediately adds: "And forget about playing with Ben tonight. Before supper, you will do nothing but your homework. Your grades are terrible lately!" Not too enthusiastically, Nathan heads towards his room to do his homework. As soon as he wants to enter it, his twin sister Olivia leaves hers, in great despair: "Nathan, you have to help me. I am on "Ben watch" tonight, but he is driving me crazy and I am expecting this really, really important phone call! Play with him until supper, will you? I'll do anything for you in return!" Nathan finds himself again in a difficult situation. He can obey his father and do the dishes. No problem there. But what should he do until supper? Olivia helps him out on a quite regular basis and he feels he ought to return the favor this time. But if he plays with Ben, he will not be able to do his homework.

This example and the one from the *preludium* have three important characteristics in common. The first is that they both concern a situation in which an agent faces several obligations, not all of which can be fulfilled. The second is that, for each of the separate obligations, there is some *prima facie* reason. In the example from the *preludium*, Nathan's specific obligations hold in view of the general rule "One ought to keep one's promises". In this last example, the obligation that Nathan ought to do the dishes holds in view of his father's command. The third characteristic is that, although not all obligations can be met, some of them can. Nathan cannot look after Ben and take Lisa to that particular movie, but he *can* go for a veggie burger in the evening. Similarly, Nathan cannot do his homework and at the same time play with Ben,

but he *can* do the dishes.

In this chapter, we will use the term *prima facie obligations* for any obligation for which there is some *prima facie* reason (some general rule, a command, ...). As the examples show (and as we all know from daily life), there are situations in which not all *prima facie* obligations can be *binding*. Nathan cannot go to that particular movie with Lisa (in view of his promise to her) and at the same time *not* go there (in view of his promise to his mother and the fact that six year olds are not allowed for this particular movie). We will use the term *actual obligations* for obligations that are binding and that should be acted upon.

Examples in which not all *prima facie* obligations can be met raise the following question: how do we decide, in a given situation, which *prima facie* obligations are actual obligations and which are not? A first answer to this question seems to be that at least those *prima facie* obligations should be considered as actual obligations that are not in conflict with any other *prima facie* obligation. This seems to capture nicely our intuitions behind the examples. The fact that Nathan made conflicting promises with respect to what he will do in the afternoon should not prevent him from going for a veggie burger in the evening. The fact that he cannot help out his twin sister as well as obey his mother should not rule out that he at least obeys his father.

This is exactly the idea behind the logics $\mathbf{SDL^r_p}$ and $\mathbf{SDL^m_p}$ presented in this section: *prima facie* obligations are considered as actual obligations *unless and until* it turns out that they are in conflict with some other *prima facie* obligation. Or, put in a somewhat different form, the logics $\mathbf{SDL^r_p}$ and $\mathbf{SDL^m_p}$ validate the inference of actual obligations from *prima facie* obligations *as much as possible*. The exact meaning of this "as much as possible" will become clear below.

The logics have two further characteristics in common: they allow us to (a) accommodate conflicts at the level of *prima facie* obligations, and (b) reason about actual obligations in the standard way (i.e., applying all axioms of **SDL**).[7] What (a) comes to is that both logics are conflict-tolerant: they do not lead to unwanted conclusions in the face of conflicting *prima facie* obligations.

We will now show, step by step, how the logics $\mathbf{SDL^r_p}$ and $\mathbf{SDL^m_p}$ are obtained.

[7] Our characteristics (a) and (b) correspond to Goble's criteria of adequacy a) and b) for *prima facie* oughts versus all-things-considered oughts [Goble, 2013, p. 257].

The lower limit logic In order to make the distinction between *prima facie* obligations and actual obligations, we will use a bi-modal language that contains two obligation operators: O^{p} and O. The first is used for *prima facie* obligations, the second for actual obligations. The language is defined as follows:

$$\mathcal{W}^{\mathsf{p}} := \mathcal{W} \mid \mathsf{O}\langle\mathcal{W}\rangle \mid \mathsf{O}^{\mathsf{p}}\langle\mathcal{W}\rangle \mid \neg\langle\mathcal{W}^{\mathsf{p}}\rangle \mid \langle\mathcal{W}^{\mathsf{p}}\rangle \vee \langle\mathcal{W}^{\mathsf{p}}\rangle \mid \langle\mathcal{W}^{\mathsf{p}}\rangle \supset \langle\mathcal{W}^{\mathsf{p}}\rangle \mid \langle\mathcal{W}^{\mathsf{p}}\rangle \wedge \langle\mathcal{W}^{\mathsf{p}}\rangle \mid \langle\mathcal{W}^{\mathsf{p}}\rangle \equiv \langle\mathcal{W}^{\mathsf{p}}\rangle$$

Note that we exclude nesting; i.e. none of the two operators occurs within the scope of another operator.

To obtain a logic that is tolerant with respect to conflicting *prima facie* obligations (characteristic (a) above), O^{p} is treated as a property-less operator, a "dummy". This means that e.g. *prima facie* obligations cannot be derived from other *prima facie* obligations. Characteristic (b) is realized by assuming that O is the ought-operator of **SDL**.

Let us call the resulting logic $\mathbf{SDL_p}$ – it is just **SDL** extended with the dummy-operator O^{p}. In AL terminology, what we have done so far is define the *lower limit logic* of our AL. This logic constitutes the monotonic core of the AL. In other words, it consists of all the principles (rules, axioms) that are unconditionally valid within the logic.[8]

In order to obtain a logic that validates the inference from *prima facie* obligations to actual obligations as much as possible, $\mathbf{SDL_p}$ needs to be strengthened. One option that does *not* work is to simply add the axiom

(A) $\mathsf{O}^{\mathsf{p}}A \supset \mathsf{O}A$

to $\mathbf{SDL_p}$. Let us call the resulting logic $\mathbf{SDL_p^+}$. In this stronger logic, conflicts at the level of *prima facie* obligations will be trivialized: if $\vdash_{\mathbf{CL}} \neg(A_1 \wedge \ldots \wedge A_n)$, then $\mathsf{O}^{\mathsf{p}}A_1, \ldots, \mathsf{O}^{\mathsf{p}}A_n \vdash_{\mathbf{SDL_p^+}} B$ for any B.[9] Of course, we could weaken the logic of O, but then we would lose characteristic (b). This shows that we need a more refined way to fulfill our aim. We will now show how this can be realized within the framework of adaptive logics.

Going adaptive What we need is a way to steer between $\mathbf{SDL_p}$ and $\mathbf{SDL_p^+}$, avoiding the weakness of the former but also the explosive

[8] The lower limit logic of every AL has to satisfy certain general desiderata, which will be spelled out in Section 3.2.

[9] To see why, note that in $\mathbf{SDL_p}$, conflicting actual obligations are trivialized just as in **SDL**. If we moreover allow for the unrestricted application of (A), this means that also conflicts at the level of *prima facie* obligations are trivialized.

character of the latter. More precisely, we need a defeasible, context-sensitive version of (A). This can be done by assuming that formulas like $\mathsf{O}^{\mathsf{P}}p \wedge \neg\mathsf{O}p$, $\mathsf{O}^{\mathsf{P}}q \wedge \neg\mathsf{O}q$, etc. are false *unless and until* proven otherwise.

In AL terminology, such formulas – the negations of defeasible assumptions – are called *abnormalities*.[10] We will use Ω_{p} to refer to the set of all those abnormalities, i.e. all formulas of the form $\mathsf{O}^{\mathsf{P}}A \wedge \neg\mathsf{O}A$.

In an adaptive proof, we can derive formulas on the assumption that certain abnormalities are false. This is most easily illustrated with an example. Let d stand for "Nathan washes the dishes", b for "Nathan plays with Ben" and h for "Nathan does his homework". The *prima facie* obligations that Nathan faces in our second running example may then be formalized as $\mathsf{O}^{\mathsf{P}}d$, $\mathsf{O}^{\mathsf{P}}b$ and $\mathsf{O}^{\mathsf{P}}(\neg b \wedge h)$. An adaptive proof from $\Gamma = \{\mathsf{O}^{\mathsf{P}}d, \mathsf{O}^{\mathsf{P}}b, \mathsf{O}^{\mathsf{P}}(\neg b \wedge h)\}$ in which we try to derive the actual obligation for Nathan to wash the dishes ($\mathsf{O}d$) may then look as follows:

1	$\mathsf{O}^{\mathsf{P}}d$	Prem	$\emptyset$
2	$\mathsf{O}^{\mathsf{P}}b$	Prem	$\emptyset$
3	$\mathsf{O}^{\mathsf{P}}(\neg b \wedge h)$	Prem	$\emptyset$
4	$\mathsf{O}d \vee \neg\mathsf{O}d$	SDL	$\emptyset$
5	$\mathsf{O}d \vee (\mathsf{O}^{\mathsf{P}}d \wedge \neg\mathsf{O}d)$	1,4; SDL	$\emptyset$
6	$\mathsf{O}d$	5; RC	$\{\mathsf{O}^{\mathsf{P}}d \wedge \neg\mathsf{O}d\}$

The fourth element of each line in this proof represents the condition of that line. This condition is always a (possibly empty) set of abnormalities. After introducing the premises on lines 1-3, we have used excluded middle to derive a new formula at line 4, and then derived line 5 using lines 1 and 4. We use "SDL" as a generic name for all rules and axioms of **SDL**. At line 6, $\mathsf{O}d$ is derived on the condition that the abnormality $\mathsf{O}^{\mathsf{P}}d \wedge \neg\mathsf{O}d$ is false. This is done by means of the rule RC (shorthand for *conditional rule*) which allows us to push abnormalities to the condition within an adaptive proof.

Here are two other applications of RC:

[10] Our terminology here and below suggests a link with Makinson's *Default Assumption Consequence Relations* [Makinson, 2005]. Indeed, as shown in [Van De Putte, 2013], one can establish an exact correspondence between Makinson's construction and ALs that use the minimal abnormality strategy.

⋮	⋮	⋮	⋮
7	$\mathsf{O}b \vee (\mathsf{O}^{\mathsf{P}}b \wedge \neg\mathsf{O}b)$	2; SDL	$\emptyset$
8	$\mathsf{O}b$	7; RC	$\{\mathsf{O}^{\mathsf{P}}b \wedge \neg\mathsf{O}b\}$
9	$\mathsf{O}(\neg b \wedge h) \vee$ $(\mathsf{O}^{\mathsf{P}}(\neg b \wedge h) \wedge \neg\mathsf{O}(\neg b \wedge h))$	3; SDL	$\emptyset$
10	$\mathsf{O}(\neg b \wedge h)$	9; RC	$\{\mathsf{O}^{\mathsf{P}}(\neg b \wedge h) \wedge$ $\neg\mathsf{O}(\neg b \wedge h)\}$

At this point, the reader may become suspicious. Clearly, $\mathsf{O}b$ and $\mathsf{O}(\neg b \wedge h)$ cannot both be true. By means of well-known **SDL**-principles, we can derive from our premises that at least one of the two corresponding abnormalities is true:

$$11 \quad (\mathsf{O}^{\mathsf{P}}b \wedge \neg\mathsf{O}b) \vee (\mathsf{O}^{\mathsf{P}}(\neg b \wedge h) \wedge \neg\mathsf{O}(\neg b \wedge h)) \quad 2,3;\ \text{SDL} \quad \emptyset$$

Formulas like the one at line 11 are called Dab-*formulas* (Dab is shorthand for "disjunction of abnormalities"). Note that this Dab-formula is derived on the empty condition. Hence, it is an unconditional consequence of the premises – it cannot be false, if the premises are true. Moreover, it is *minimal*: neither of its disjuncts $\mathsf{O}^{\mathsf{P}}b \wedge \neg\mathsf{O}b$ or $\mathsf{O}^{\mathsf{P}}(\neg b \wedge h) \wedge \neg\mathsf{O}(\neg b \wedge h)$ is derived on the empty condition in the above proof.[11]

At lines 8 and 10 respectively, we relied on the assumption that the first, respectively the second of these abnormalities is false. But line 11 clearly indicates that those two assumptions cannot be jointly true. So a mechanism is needed to *retract* the inferences at lines 8 and 10.

Formally, this is taken care of by a *marking definition*, which stipulates which lines are marked, and hence considered "out" at a given stage of an adaptive proof. How the marking proceeds depends on the so-called *adaptive strategy*. The logics $\mathbf{SDL}^{\mathbf{r}}_{\mathbf{p}}$ and $\mathbf{SDL}^{\mathbf{m}}_{\mathbf{p}}$ are based respectively based on the Reliability strategy and the Minimal Abnormality strategy. Let us look at these in turn.

Reliability For $\mathbf{SDL}^{\mathbf{r}}_{\mathbf{p}}$, a line is marked whenever its condition contains an abnormality that is a disjunct of a minimal Dab-formula that has been derived in the same proof. For instance, in the above example, lines 8 and 10 are marked, whereas all other lines are not marked. This is indicated by a ✓-symbol at the end of the line:

[11] In fact, neither of them *can* be derived in this proof on the empty condition, since they simply do not follow from Γ by $\mathbf{SDL}_{\mathbf{p}}$.

1	$\mathsf{O}^{\mathsf{P}}d$	Prem	$\emptyset$
2	$\mathsf{O}^{\mathsf{P}}b$	Prem	$\emptyset$
3	$\mathsf{O}^{\mathsf{P}}(\neg b \wedge h)$	Prem	$\emptyset$
4	$\mathsf{O}d \vee \neg\mathsf{O}d$	SDL	$\emptyset$
5	$\mathsf{O}d \vee (\mathsf{O}^{\mathsf{P}}d \wedge \neg\mathsf{O}d)$	1,4; SDL	$\emptyset$
6	$\mathsf{O}d$	5; RC	$\{\mathsf{O}^{\mathsf{P}}d \wedge \neg\mathsf{O}d\}$
7	$\mathsf{O}b \vee (\mathsf{O}^{\mathsf{P}}b \wedge \neg\mathsf{O}b)$	2; SDL	$\emptyset$
8	$\mathsf{O}b$	7; RC	$\{\mathsf{O}^{\mathsf{P}}b \wedge \neg\mathsf{O}b\}$ ✓
9	$\mathsf{O}(\neg b \wedge h) \vee$ $(\mathsf{O}^{\mathsf{P}}(\neg b \wedge h) \wedge \neg\mathsf{O}(\neg b \wedge h))$	3; SDL	$\emptyset$
10	$\mathsf{O}(\neg b \wedge h)$	9; RC	$\{\mathsf{O}^{\mathsf{P}}(\neg b \wedge h) \wedge$ $\neg\mathsf{O}(\neg b \wedge h)\}$ ✓
11	$(\mathsf{O}^{\mathsf{P}}b \wedge \neg\mathsf{O}b) \vee$ $(\mathsf{O}^{\mathsf{P}}(\neg b \wedge h) \wedge \neg\mathsf{O}(\neg b \wedge h))$	2,3; SDL	$\emptyset$

In general, lines with an empty condition are never marked. But also those lines whose condition is not problematic in view of the minimal Dab-formulas in the proof remain unmarked (witness line 6 in the example). So at the end of the day, some instances of (A) are trustworthy in the light of the premises, while other instances of (A) are not. This illustrates the premise-sensitivity of adaptive logics that was mentioned in Section 1.[12]

The fact that lines can become marked in a proof means that we cannot simply define logical consequence in terms of being derivable in a proof. We need a more robust notion of derivability; this is called *final derivability*. The basic idea is that something is finally derivable if and only if it can be derived in a "stable" way. Spelling out this intuition is not as straightforward as it may seem, as it requires quantification over extensions of proofs. We refer to Definitions 3.3 and 3.4 in the next section for the exact details.

Minimal Abnormality The logic $\mathbf{SDL}^{\mathbf{m}}_{\mathbf{p}}$ works in exactly the same way as $\mathbf{SDL}^{\mathbf{r}}_{\mathbf{p}}$, except that the marking in both logics is slightly different. Consider the following extension of our proof:

[12]Some may argue that, in light of the premise set, the inferences at lines 8 and 10 were never rational in the first place. Admittedly, in cases like Γ above, it can easily be seen which *prima facie* obligations can make it into actual obligations, and which cannot on pain of triviality. But then again, such cases are not the only ones we may encounter in practice. Conflicts may exist between many different *prima facie* obligations, and they may be very hard to trace. Once we move to the predicate level, it may even be undecidable whether a certain set of *prima facie* obligations is consistent. One may well be calculating up to eternity before ever knowing for sure whether a certain inference is safe.

1	$\mathsf{O}^{\mathsf{P}}d$	Prem	$\emptyset$
2	$\mathsf{O}^{\mathsf{P}}b$	Prem	$\emptyset$
3	$\mathsf{O}^{\mathsf{P}}(\neg b \wedge h)$	Prem	$\emptyset$
4	$\mathsf{O}d \vee \neg\mathsf{O}d$	SDL	$\emptyset$
5	$\mathsf{O}d \vee (\mathsf{O}^{\mathsf{P}}d \wedge \neg\mathsf{O}d)$	1,4; SDL	$\emptyset$
6	$\mathsf{O}d$	5; RC	$\{\mathsf{O}^{\mathsf{P}}d \wedge \neg\mathsf{O}d\}$
7	$\mathsf{O}b \vee (\mathsf{O}^{\mathsf{P}}b \wedge \neg\mathsf{O}b)$	2; SDL	$\emptyset$
8	$\mathsf{O}b$	7; RC	$\{\mathsf{O}^{\mathsf{P}}\neg r \wedge \neg\mathsf{O}\neg r\}$ ✓
9	$\mathsf{O}(\neg b \wedge h) \vee$ $(\mathsf{O}^{\mathsf{P}}(\neg b \wedge h) \wedge \neg\mathsf{O}(\neg b \wedge h))$	3; SDL	$\emptyset$
10	$\mathsf{O}(\neg b \wedge h)$	9; RC	$\{\mathsf{O}^{\mathsf{P}}(\neg b \wedge h) \wedge$ $\neg\mathsf{O}(\neg b \wedge h)\}$ ✓
11	$(\mathsf{O}^{\mathsf{P}}b \wedge \neg\mathsf{O}b) \vee$ $(\mathsf{O}^{\mathsf{P}}(\neg b \wedge h) \wedge \neg\mathsf{O}(\neg b \wedge h))$	2,3; SDL	$\emptyset$
12	$\mathsf{O}(b \vee h)$	8; (Inh)	$\{\mathsf{O}^{\mathsf{P}}b \wedge \neg\mathsf{O}b\}$?
13	$\mathsf{O}(b \vee h)$	10; (Inh)	$\{\mathsf{O}^{\mathsf{P}}(\neg b \wedge h) \wedge$ $\neg\mathsf{O}(\neg b \wedge h)\}$?

Note first that, since we used the formula at line 8 to derive the one at line 12, the latter inherits the condition of the former. Likewise, line 13 is derived on the same condition as line 10. Taken together, lines 12 and 13 indicate that $\mathsf{O}(b \vee h)$ is true if either of the abnormalities in the Dab-formula at line 11 is false.

Should lines 12 and 13 in this proof be marked? Clearly, there is a problem with at least one of the two involved abnormalities. Since there is no reason to prefer the falsehood of one over that of the other, that means both abnormalities are "unreliable" at this proof stage. However, if we assume that as few abnormalities as possible are true – until and unless proven otherwise –, then in cases like these we will assume that only one of both abnormalities is true. And in that case, $\mathsf{O}(b \vee h)$ does follow.

To turn this idea into a general method for marking lines in an adaptive proof, we need the concept of a ($\subset$-minimal) choice set. Suppose that the Dab-formulas at the current stage of our proof are $\mathsf{Dab}(\Delta_1)$, $\mathsf{Dab}(\Delta_2), \ldots$. A choice set of $\{\Delta_1, \Delta_2, \ldots\}$ is a set φ that contains at least one member of each Δ_i. In view of our proof, we know that (at least) the members of one choice set of $\{\Delta_1, \Delta_2, \ldots\}$ should be true in view of the premises. However, we are still free to assume that *only* the members of a $\subset$-minimal choice set of $\{\Delta_1, \Delta_2, \ldots\}$ are true. Suppose now moreover that, for every such minimal choice set φ, we can derive

A on a condition Θ that does not overlap with φ. This means that we have sufficient reasons to infer A – since every minimally abnormal way of interpreting the current proof stage will make A true. Following this general line of reasoning, lines 12 and 13 will not be marked, but lines 8 and 10 will be marked just as before.

To summarize: one can be cautious to different degrees when reasoning defeasibly; this difference is modeled by the adaptive strategy. According to the *reliability strategy* (usually indicated with a superscript r), both lines 12 and 13 are marked. According to *minimal abormality*, they are both unmarked. In general, reliability is slightly weaker (more cautious) than minimal abnormality – see Theorem 3.15.

We now turn to the general characterization of ALs. A critical discussion of the logics $\mathbf{SDL}^{\mathbf{r}}_{\mathbf{p}}$ and $\mathbf{SDL}^{\mathbf{m}}_{\mathbf{p}}$ is postponed until Section 8. There we evaluate $\mathbf{SDL}^{\mathbf{r}}_{\mathbf{p}}$ and $\mathbf{SDL}^{\mathbf{m}}_{\mathbf{p}}$ by various criteria that are introduced in Section 4.

3.2 The standard format

The locus classicus for the standard format is Batens' [2007]; an earlier version of it appeared in [Batens, 2001]. Here, we will follow the more recent presentation from [Batens, 2015], indicating minor differences where they occur. We will only explain the general characteristics, and refer to the works just cited for more details.

Standardly, a logic is defined as a function $\mathbf{L} : \wp(\mathcal{W}_{\mathbf{L}}) \rightarrow \wp(\mathcal{W}_{\mathbf{L}})$, where $\mathcal{W}_{\mathbf{L}}$ is the set of formulas in the formal language of $\mathbf{L}$. This also holds for adaptive logics. For adaptive logics in standard format, the language should at least contain the classical disjunction $\vee$.[13] For reasons of convenience, we will in this chapter assume that the language also contains the classical negation $\neg$.

Every logic $\mathbf{AL}^{\mathbf{x}}$ is defined by a triple:

1. A lower limit logic **LLL**. This is a reflexive, transitive, monotonic and compact logic[14] that has a characteristic semantics and for which at least the disjunction $\vee$ behaves classically.

[13] The assumption that the language contains a classical disjunction can be questioned on philosophical grounds. In [Odintsov and Speranski, 2013; Batens, 2015] it is shown that one can do without this assumption, if one rephrases everything in terms of multi-conclusion sequents.

[14] Let Cn be the consequence operation of a logic $\mathbf{L}$. $\mathbf{L}$ is *reflexive* iff for all Γ, $\Gamma \subseteq \mathsf{Cn}_{\mathbf{L}}(\Gamma)$. $\mathbf{L}$ is *transitive* iff for all Γ, Γ': if $\Gamma' \subseteq \mathsf{Cn}_{\mathbf{L}}(\Gamma)$, then $\mathsf{Cn}_{\mathbf{L}}(\Gamma \cup \Gamma') \subseteq \mathsf{Cn}_{\mathbf{L}}(\Gamma)$. $\mathbf{L}$ is *monotonic* iff for all Γ, Γ', $\mathsf{Cn}_{\mathbf{L}}(\Gamma) \subseteq \mathsf{Cn}_{\mathbf{L}}(\Gamma \cup \Gamma')$. $\mathbf{L}$ is *compact* iff for all Γ, A, if $A \in \mathsf{Cn}_{\mathbf{L}}(\Gamma)$, then there is a finite $\Gamma' \subseteq \Gamma$ with $A \in \mathsf{Cn}_{\mathbf{L}}(\Gamma')$.

2. A set of abnormalities $\Omega \subseteq \mathcal{W}_{\mathbf{LLL}}$ that is specified in terms of one or several logical forms.
3. An adaptive strategy: Reliability (when $\mathsf{x} = \mathsf{r}$) or Minimal Abnormality (when $\mathsf{x} = \mathsf{m}$).

For instance, the adaptive logic $\mathbf{SDL}^{\mathbf{r}}_{\mathbf{p}}$ from Section 3.1 is defined by the triple $\langle \mathbf{SDL_p}, \Omega_{\mathsf{p}}, \mathsf{r} \rangle$; the logic $\mathbf{SDL}^{\mathbf{m}}_{\mathbf{p}}$ is defined by $\langle \mathbf{SDL_p}, \Omega_{\mathsf{p}}, \mathsf{m} \rangle$. The logical form that specifies Ω_{p} is $\mathsf{O}^{\mathsf{p}}A \wedge \neg \mathsf{O}A$. In general, it is required that only countably many logical forms specify the set of abnormalities.

In the remainder of this section, we presuppose a fixed **LLL**, Ω, and strategy $\mathsf{x} \in \{\mathsf{r}, \mathsf{m}\}$. We use $\mathsf{Dab}(\Delta)$ to denote the (classical) disjunction of the members of Δ, where it is presupposed that Δ is a finite subset of Ω.

Proof theory The core idea behind the adaptive proof theory is to take all the inference rules of the lower limit logic for granted and to allow in addition for defeasible applications of some rules. Defeasible inferences in adaptive proofs are conditional. Hence, the usual way in which lines in proofs are presented – by a line number, a formula, and a justification – is enriched by a fourth element: a condition. A condition in turn is a set of abnormalities.

Suppose some formula A is derived on the condition $\{B_1, B_2, \ldots, B_n\} \subseteq \Omega$. The intended reading is that A is derived on the assumption that all the abnormalities $B_1, \ldots, B_n$ are false.

Adaptive proofs are characterized by three generic rules and a marking definition. Let us first discuss the generic rules. In what follows we skip the line numbers and justification of lines.

Prem If $A \in \Gamma$:

$$\frac{\begin{array}{cc} \vdots & \vdots \end{array}}{\begin{array}{cc} A & \emptyset \end{array}}$$

RU If $A_1, \ldots, A_n \vdash_{\mathbf{LLL}} B$:

$$\frac{\begin{array}{cc} A_1 & \Delta_1 \\ \vdots & \vdots \\ A_n & \Delta_n \end{array}}{\begin{array}{cc} B & \Delta_1 \cup \ldots \cup \Delta_n \end{array}}$$

RC If $A_1, \ldots, A_n \vdash_{\mathbf{LLL}} B \vee \mathsf{Dab}(\Theta)$:

$$\frac{\begin{array}{cc} A_1 & \Delta_1 \\ \vdots & \vdots \\ A_n & \Delta_n \end{array}}{\begin{array}{cc} B & \Delta_1 \cup \ldots \cup \Delta_n \cup \Theta \end{array}}$$

By means of Prem, any premise may be introduced on the empty

condition. Of course, we do not need any defeasible assumptions in order to state premises. The unconditional rule (RU) makes it possible to apply any inference rule of **LLL** in an adaptive proof. Note that RU may also be applied to lines that were derived on defeasible assumptions, i.e. where $\Delta_i \neq \emptyset$ for some $i \in \{1, \ldots, n\}$. The assumptions under which the A_i's were derived thus carry forward to the line at which B is derived. In virtue of Prem and RU, ALs inherit all the inferential power of **LLL**: any **LLL**-proof can be rephrased as an **AL**-proof just by adding the empty condition in the fourth column and by replacing the respective **LLL**-rules by Prem or RU.

In Section 3.1, we sometimes referred explicitly to the axiom that was used to derive a specific line in an adaptive proof. In the remainder we use RU as a metavariable for all axioms and (derivable) rules of the **LLL**; whenever useful we will indicate in footnotes which exact axioms were applied in order to derive a new line.

The rule that permits the introduction of new conditions in an adaptive proof is RC, the conditional rule. Suppose that we can derive $B \vee \mathsf{Dab}(\Theta)$ by means of **LLL**, i.e. that either B is the case or some of the abnormalities in Θ. Then RC allows us to derive B on the assumption that none of the abnormalities in Θ is true. Making this assumption amounts to adding all members of Θ to the condition by means of RC. Similarly as for RU, in case some of the lines that are used for the inference step are conditional inferences, we carry forward their conditions as well.

Apart from the possibility to make conditional derivations via RC, a second distinctive aspect of adaptive proofs is the marking definition, which is applied at each *stage* of a proof. A stage is simply a sequence of lines, obtained by the application of the above rules. For concrete examples, we will identify stages with their last line. So for example the last stage of the last proof displayed in Section 3.1 is referred to as stage 13.

$\mathsf{Dab}(\Delta)$ is a Dab-formula at stage s of a proof, iff it is the second element of a line of the proof with an empty condition, and derived by means of RU.[15] $\mathsf{Dab}(\Delta)$ is a *minimal* Dab-formula at stage s iff there is no other Dab-formula $\mathsf{Dab}(\Delta')$ at stage s such that $\Delta' \subset \Delta$. Where $\mathsf{Dab}(\Delta_1), \mathsf{Dab}(\Delta_2), \ldots$ are the minimal Dab-formulas at stage s of a proof, let $\Sigma_s(\Gamma) = \{\Delta_1, \Delta_2, \ldots\}$. Finally, let $U_s(\Gamma) = \bigcup \Sigma_s(\Gamma)$.

[15] Here, our terminology differs slightly from that in [Batens, 2015]. Batens uses the term "Dab-formula at stage s" for any disjunction of abnormalities derived at s, whereas we preserve it for those that have been derived by means of RU. Batens calls the latter "*inferred* Dab-formulas".

Definition 3.1 (Marking for $\mathbf{AL^r}$). *A line l is marked at stage s iff, where Δ is its condition, $\Delta \cap U_s(\Gamma) \neq \emptyset$.*

In terms of assumptions, this means that according to the reliability strategy, an assumption is "safe" at stage s iff the corresponding abnormality is not a member of $U_s(\Gamma)$, and an inference is "safe" at s iff it only relies on assumptions that are safe at s.

Returning to our example of page 382, we can see that $\Sigma_{11}(\Gamma) = \{\{\mathsf{O^P}b \wedge \neg \mathsf{O}b, \mathsf{O^P}(\neg b \wedge h) \wedge \neg \mathsf{O}(\neg b \wedge h)\}\}$ and hence $U_{11}(\Gamma) = \{\mathsf{O^P}b \wedge \neg \mathsf{O}b, \mathsf{O^P}(\neg b \wedge h) \wedge \neg \mathsf{O}(\neg b \wedge h)\}$. This explains why lines 8 and 10 are marked at stage 11 of the proof.

The marking definition for minimal abnormality requires some more terminology. Recall that, where Σ is a set of sets, φ is a *choice set* of Σ iff for every $\Delta \in \Sigma$, $\varphi \cap \Delta \neq \emptyset$. φ is a *minimal* choice set of Σ iff there is no choice set ψ of Σ such that $\psi \subset \varphi$. Let $\Phi_s(\Gamma)$ be the set of $\subset$-minimal choice sets of $\Sigma_s(\Gamma)$. Marking for minimal abnormality proceeds as follows:

Definition 3.2 (Marking for $\mathbf{AL^m}$). *A line l with formula A is marked at stage s iff, where its condition is Δ: (i) there is no $\varphi \in \Phi_s(\Gamma)$ such that $\varphi \cap \Delta = \emptyset$, or (ii) for a $\varphi \in \Phi_s(\Gamma)$, there is no line at which A is derived on a condition Θ for which $\Theta \cap \varphi = \emptyset$.*

In our simple example on page 384, $\Phi_{13}(\Gamma) = \{\{\mathsf{O^P}b \wedge \neg \mathsf{O}b\}, \{\mathsf{O^P}(\neg b \wedge h) \wedge \neg \mathsf{O}(\neg b \wedge h)\}\}$. In view of condition (ii) in Definition 3.2, lines 8 and 10 are marked for minimal abnormality at stage 13, but lines 12 and 13 are not. Note that all of these lines are marked for reliability.

If a line that has A as its second element is marked at stage s, this indicates that according to our best insights at this stage, A cannot be considered derivable. If the line is unmarked at stage s, we say that A is derivable at stage s of the proof. Since marks may come and go as a proof proceeds, we also need to define a stable notion of derivability. This definition is the same for both strategies.

Where s is a proof stage, an *extension* of s is every stage s' that contains the lines occurring in s in the same order. Hence putting lines in front of s, inserting them somewhere in between lines of s, or simply adding them at the end of s may all result in an extension of s.

Definition 3.3. *A is* finally derived *from Γ at line l of a stage s iff (i) A is the second element of line l, (ii) line l is unmarked at s, and (iii) every extension of s in which line l is marked may be further extended in such a way that line l is unmarked again.*

Definition 3.4. $\Gamma \vdash_{\mathbf{AL^x}} A$ *($A \in \mathsf{Cn}_{\mathbf{AL^x}}(\Gamma)$) iff A is finally derived at a line of a stage in an* $\mathbf{AL^x}$*-proof from* Γ.

Note that in order to be finally derivable, A must be derived at a line l, where $l \in \mathbb{N}$. This means that every formula that is finally derivable from Γ can be finally derived in a *finite* proof from Γ. However, we need a meta-level argument to show that clauses (ii) and (iii) in Definition 3.3 are satisfied, and hence that $\Gamma \vdash_{\mathbf{AL^x}} A$.

Semantics On the supposition that **LLL** is characterized by a model theoretic semantics (with the semantic consequence relation $\Vdash_{\mathbf{LLL}}$), one can also give a semantics for $\mathbf{AL^x}$. The rough idea is as follows: from the set of **LLL**-models of a given premise set, $\mathbf{AL^x}$ selects a subset of "preferred" models. Whatever holds in those preferred models, follows by $\mathbf{AL^x}$.[16]

What counts as a preferred model depends on the strategy used. For minimal abnormality, only those models of the premise set are selected which verify a $\subset$-minimal set of abnormalities. For reliability, a threshold of *unreliable* abnormalities (with respect to a given premise set Γ) is defined, and only the models that do not verify any abnormalities other than the unreliable ones, are selected.

To define the $\mathbf{AL^x}$-semantics in exact terms, we need some more notation. Validity of a formula A in a model M will be written as $M \models A$. M is an **LLL**-model of Γ iff $M \models A$ for all $A \in \Gamma$. $\mathcal{M}_{\mathbf{LLL}}(\Gamma)$ denotes the set of **LLL**-models of Γ. Where M is an **LLL**-model, its *abnormal part* is given by $\mathsf{Ab}(M) =_{\mathsf{df}} \{B \mid B \in \Omega, M \models B\}$.

For reliability, the selection of preferred models is in some sense analogous to the marking definition. $\mathsf{Dab}(\Delta)$ is a *minimal* Dab*-consequence of* Γ iff $\Gamma \Vdash_{\mathbf{LLL}} \mathsf{Dab}(\Delta)$ and there is no $\Delta' \subset \Delta$ for which $\Gamma \Vdash_{\mathbf{LLL}} \mathsf{Dab}(\Delta')$. Where $\mathsf{Dab}(\Delta_1), \mathsf{Dab}(\Delta_2), \ldots$ are the minimal Dab-consequences of Γ, let $\Sigma(\Gamma) = \{\Delta_1, \Delta_2, \ldots\}$. Let $U(\Gamma) = \bigcup \Sigma(\Gamma)$. We say that $U(\Gamma)$ is the set of *unreliable* formulas with respect to Γ.

Definition 3.5. *An* **LLL**-*model M of Γ is* reliable *iff* $\mathsf{Ab}(M) \subseteq U(\Gamma)$.

Definition 3.6. $\Gamma \Vdash_{\mathbf{AL^r}} A$ *iff A is verified by all reliable models of* Γ.

[16] Note that this is similar to the semantics of circumscription (where models are selected in which the abnormal predicates have a minimal extension) and Shoham-style preferential semantics (where all the $\prec$-minimal models are selected, for a given order $\prec$ on the models of a premise set). However, in ALs, the selection depends on purely syntactic properties of the models, viz. the formulas (more specifically, the abnormalities) that they verify. This in turn gives ALs fairly strong meta-theoretic properties – see Section 3.3.

For minimal abnormality, the semantics' simplicity stands in sharp contrast to the intricate marking definition:

Definition 3.7. *An* **LLL***-model* M *of* Γ *is* minimally abnormal *iff there is no* **LLL***-model* M' *of* Γ *such that* $\mathsf{Ab}(M') \subset \mathsf{Ab}(M)$.

Definition 3.8. $\Gamma \Vdash_{\mathbf{AL^m}} A$ *iff* A *is verified by all minimally abnormal models of* Γ.

In the remainder, we will denote the set of $\mathbf{AL^x}$-models of a set Γ by $\mathcal{M}_{\mathbf{AL^x}}(\Gamma)$.

Upper Limit Logic The so-called *upper limit logic* of $\mathbf{AL^x}$ is defined as the Tarski-logic[17] obtained by adding all negations of abnormalities as axioms to **LLL**. That is, where $\Omega^{\neg} = \{\neg A \mid A \in \Omega\}$, $\Gamma \vdash_{\mathbf{ULL}} A$ iff $\Gamma \cup \Omega^{\neg} \vdash_{\mathbf{LLL}} A$. By the compactness of **LLL**, $\Gamma \vdash_{\mathbf{ULL}} A$ iff there are $B_1, \ldots, B_n \in \Omega$ such that $\Gamma \cup \{\neg B_1, \ldots, \neg B_n\} \vdash_{\mathbf{LLL}} A$. $\mathbf{AL^x}$ can be seen as steering a middle course between **LLL** and **ULL** (see Theorem 3.15 below).

In our running example, $\mathbf{SDL^+_p}$ is the upper limit logic of both $\mathbf{SDL^r_p}$ and $\mathbf{SDL^m_p}$. Note that in general, **ULL** does not depend on the strategy of $\mathbf{AL^x}$.

3.3 Some meta-properties of ALs in standard format

Once defined within the standard format, it is guaranteed that an AL satisfies a number of meta-properties. We only mention some of them here for the ease of reference. Their proofs can be found in [Batens, 2007].

First of all, the dynamic proof theory is sound and complete with respect to the semantics of $\mathbf{AL^x}$:

Theorem 3.9 (Soundness and Completeness)**.** $\Gamma \vdash_{\mathbf{AL^x}} A$ *iff* $\Gamma \Vdash_{\mathbf{AL^x}} A$.

It follows from this result that one can rely on semantic considerations in order to prove that a formula A is finally derivable from a given Γ. We will in the remainder rely freely on Theorem 3.9, switching between the semantic and proof theoretic consequence relation where suitable.

Recall that the semantics of an AL consists in selecting a subset of the **LLL**-models of Γ. Now, when a model M is not selected, we should

[17] A Tarski-logic is a logic whose consequence relation is reflexive, monotonic, and transitive.

be able to justify this in terms of another model M' that *is* selected, and is more normal than M. This is what the following theorem gives us:

Theorem 3.10 (Strong Reassurance)**.** *If* $M \in \mathcal{M}_{\mathbf{LLL}}(\Gamma) - \mathcal{M}_{\mathbf{AL^x}}(\Gamma)$*, then there is an* $M' \in \mathcal{M}_{\mathbf{AL^x}}(\Gamma)$ *such that* $\mathsf{Ab}(M') \subset \mathsf{Ab}(M)$*.*

In other words, the preference relation defined in terms of $\subset$ and the abnormal part relation is smooth with respect to every set $\mathcal{M}_{\mathbf{LLL}}(\Gamma)$.[18] It is well-known that a selection semantics based on such a smooth preference relation warrants the following properties in turn:[19]

Theorem 3.11 (Consistency Preservation)**.** *If* Γ *has* **LLL***-models, then* $\mathcal{M}_{\mathbf{AL^x}}(\Gamma) \neq \emptyset$*. Hence,* Γ *is* $\mathbf{AL^x}$*-trivial iff* Γ *is* **LLL***-trivial.*

Theorem 3.12 (Cumulative Indifference)**.** *If* $\Gamma' \subseteq \mathsf{Cn}_{\mathbf{AL^x}}(\Gamma)$*, then* $\mathsf{Cn}_{\mathbf{AL^x}}(\Gamma) = \mathsf{Cn}_{\mathbf{AL^x}}(\Gamma \cup \Gamma')$*.*

In the literature on non-monotonic logics, cumulative indifference is often divided into two properties: cumulative transitivity or cut (if $\Gamma' \subseteq \mathsf{Cn}_{\mathbf{AL^x}}(\Gamma)$, then $\mathsf{Cn}_{\mathbf{AL^x}}(\Gamma \cup \Gamma') \subseteq \mathsf{Cn}_{\mathbf{AL^x}}(\Gamma)$) and cumulative or cautious monotonicity (if $\Gamma' \subseteq \mathsf{Cn}_{\mathbf{AL^x}}(\Gamma)$, then $\mathsf{Cn}_{\mathbf{AL^x}}(\Gamma) \subseteq \mathsf{Cn}_{\mathbf{AL^x}}(\Gamma \cup \Gamma')$).

Strong reassurance, consistency preservation, and cumulative indifference are generally considered desirable for non-monotonic consequence relations, see e.g. [Makinson, 2005]. It speaks in favor of ALs (in standard format) that they satisfy each of these properties. In particular, cautious monotonicity is a very intuitive property: if a formula follows from a premise set Γ, then it ought to follow from any Γ' that is obtained by extending Γ with some logical consequences of Γ. The extended premise set Γ' contains no genuinely new information, as the additions are in a sense already contained in Γ.

Suppose that Γ and Γ' are **LLL**-equivalent, i.e. $\mathsf{Cn}_{\mathbf{LLL}}(\Gamma) = \mathsf{Cn}_{\mathbf{LLL}}(\Gamma')$. It follows that they have the same set of **LLL**-models and that $U(\Gamma) = U(\Gamma')$. Hence in view of the semantics, they will also have the same $\mathbf{AL^x}$-models, and hence be $\mathbf{AL^x}$-equivalent. So we have a fairly straightforward criterion to decide when two premise sets are equivalent according to $\mathbf{AL^x}$:[20]

Theorem 3.13 (Equivalence)**.** *If* $\mathsf{Cn}_{\mathbf{LLL}}(\Gamma) = \mathsf{Cn}_{\mathbf{LLL}}(\Gamma')$*, then* $\mathsf{Cn}_{\mathbf{AL^x}}(\Gamma) = \mathsf{Cn}_{\mathbf{AL^x}}(\Gamma')$*.*

[18] A partial order $\prec$ is *smooth* with respect to a set X iff for all $x \in X$, either x is $\prec$-minimal in X, or there is some $\prec$-minimal $y \in X$ such that $y \prec x$.

[19] See e.g. [Makinson, 1994].

[20] Similar criteria for equivalence are discussed in [Batens *et al.*, 2009]; an extended and updated version of this paper can be found in [Straßer, 2014, Chapter 4].

The next property on the list is specific to ALs, as it concerns the notion of an abnormality. It will be of particular use in Sections 5-7.

Say a premise set Γ is *normal* iff $\Gamma \cup \{\neg A \mid A \in \Omega\}$ is not **LLL**-trivial; in other words, iff it is **ULL**-consistent. The theorem states that every adaptive logic is as powerful as its upper limit logic when normal premise sets are concerned:[21]

Theorem 3.14 (**ULL**-recapture)**.** Γ *is a normal premise set iff* $\mathsf{Cn}_{\mathbf{AL}^{\mathsf{x}}}(\Gamma) = \mathsf{Cn}_{\mathbf{ULL}}(\Gamma)$.

The last theorem simply recalls the relation between **LLL**, $\mathbf{AL}^{\mathbf{r}}$, $\mathbf{AL}^{\mathbf{m}}$ and **ULL**, which was illustrated in Section 3.1:

Theorem 3.15. $\mathsf{Cn}_{\mathbf{LLL}}(\Gamma) \subseteq \mathsf{Cn}_{\mathbf{AL}^{\mathbf{r}}}(\Gamma) \subseteq \mathsf{Cn}_{\mathbf{AL}^{\mathbf{m}}}(\Gamma) \subseteq \mathsf{Cn}_{\mathbf{ULL}}(\Gamma)$.

3.4 Variants and extensions of the standard format

In this section, we briefly consider two variants of the standard format that are useful in the context of deontic reasoning; we will occasionally refer back to both variants in the remainder of this chapter. We focus on the essential ideas in both cases; the metatheory of these (and many other) variants of the standard format is studied at length in [Straßer, 2014, Chapter 5].

Normal Selections The minimal abnormality strategy corresponds to what is called the *skeptical* solution to the problem of multiple extensions in default logic.[22] That is, A is finally $\mathbf{AL}^{\mathbf{m}}$-derivable from Γ if and only if, for *every* maximal set $\Delta \subseteq \Omega^{\neg}$ such that $\Gamma \cup \Delta$ is **LLL**-satisfiable, $\Gamma \cup \Delta \vdash_{\mathbf{LLL}} A$.[23]

Rather than taking the universal quantification over such maximal sets, one may also quantify existentially over them. That is, say $\Gamma \vdash_{\mathbf{AL}^{\mathsf{n}}} A$ iff *there is a* maximal set $\Delta \subseteq \Omega^{\neg}$ such that $\Gamma \cup \Delta \vdash_{\mathbf{LLL}} A$. The superscript n refers to "normal selections", which is the name of the adaptive strategy of the resulting logics. Proof-theoretically, such logics are characterized in exactly the same way as ALs in standard format, with the only exception that the marking definition is simplified:

[21] Our name for the theorem is inspired by discussions in paraconsistent logic, where a similar property is called "classical recapture" [Priest, 1987].

[22] Analogous problems arise in Input/Output-logic, inheritance networks, and abstract argumentation, giving rise to similar distinctions between less and more cautious "modes of reasoning" – see [Straßer, 2014, Sect. 2.8] for more discussion.

[23] This is a well-known property that is often used in the metatheory of ALs; see e.g. [Van De Putte, 2013] for a proof of it.

Definition 3.16 (Marking for Normal Selections)**.** *A line l in a proof with condition Δ is marked at stage s iff* $\mathsf{Dab}(\Delta)$ *is derived on the empty condition at s.*

The consequence relation $\vdash_{\mathbf{AL^n}}$ is usually very strong, and yet does not trivialize premise sets as long as they are **LLL**-consistent. However, it will not in general be closed under **LLL**. More generally, many of the nice properties we discussed in Section 3.3 can fail for $\vdash_{\mathbf{AL^n}}$.

To understand this, consider the logic $\mathbf{SDL^n_p}$, defined by the triple $\langle \mathbf{SDL_p}, \Omega_\mathbf{p}, \text{normal selections} \rangle$. Let $\Gamma = \{\mathsf{O}^\mathsf{P}p, \mathsf{O}^\mathsf{P}q, \neg\mathsf{O}(p \wedge q)\}$. Note that this premise set has the following minimal Dab-consequence:

$$(\mathsf{O}^\mathsf{P}p \wedge \neg\mathsf{O}p) \vee (\mathsf{O}^\mathsf{P}q \wedge \neg\mathsf{O}q) \tag{1}$$

Since this is a minimal Dab-consequence of Γ, both $\mathsf{O}p$ and $\mathsf{O}q$ are individually compatible with Γ. Hence, both $\mathsf{O}p$ and $\mathsf{O}q$ are finally $\mathbf{SDL^n_p}$-derivable from Γ, on the respective conditions $\{\mathsf{O}^\mathsf{P}p \wedge \neg\mathsf{O}p\}$ and $\{\mathsf{O}^\mathsf{P}q \wedge \neg\mathsf{O}q\}$. However, $\mathsf{O}p \wedge \mathsf{O}q$ is not finally $\mathbf{SDL^n_p}$-derivable from Γ, since one needs to rely on the falsity of both abnormalities in order to obtain this conclusion. This shows that the consequence relation of $\mathbf{SDL^n_p}$ is not closed under the rule of conjunction, even if $\wedge$ behaves classically in the lower limit logic.

In the context of deontic logic, normal selections has been used to characterize one variant of Horty's approach to conflicting obligations [Straßer *et al.*, 2017]. Likewise, it has been applied to characterize constrained Input/Output-logics that are defined in terms of the join of the maximal unconflicted sets of generators [Straßer *et al.*, 2016]. We will shortly return to the latter systems in Section 9.2.

Prioritized adaptive logics Another useful variation of the standard format is obtained by distinguishing between various types of abnormalities, and by giving priority to some of these when minimizing abnormality. This can be done in at least three clearly distinct ways – see [Van De Putte, 2012] for a detailed study of these. Here we will only discuss one of these three, viz. the so-called *lexicographic adaptive logics* first presented in [Van De Putte and Straßer, 2012]; we moreover confine ourselves to the minimal abnormality-variant of these systems. Although these logics can be fully characterized in terms of a dynamic proof theory, we focus on their semantics, which is a straightforward generalization of the $\mathbf{AL^m}$-semantics.

Let $\langle \Omega_i \rangle_{i \in I}$ (for $I \subseteq \mathbb{N}$) be a sequence of sets of abnormalities. Intuitively, the idea is that we consider the members of Ω_1 to be the "worst"

abnormalities; those of Ω_2 as "slightly less problematic (yet still abnormal)", etc. Thus, we want to make sure when selecting models, that we first minimize with respect to Ω_1, next with respect to Ω_2, etc. This is done in terms of a lexicographic order $\sqsubset$ on the abnormal parts of the models:

Definition 3.17. *Where* $\Delta, \Delta' \subseteq \bigcup_{i \in I} \Omega_i$*:* $\Delta \sqsubset \Delta'$ *iff there is a* $j \in I$ *such that (1) for all* $k < j$ *(if any),* $\Delta \cap \Omega_k = \Delta' \cap \Omega_k$ *and (2)* $\Delta \cap \Omega_j \subset \Delta' \cap \Omega_j$*.*

The preference relation $\sqsubset$ on abnormal parts of models yields a smooth preference relation on every set $\mathcal{M}_{\mathbf{LLL}}(\Gamma)$ [Van De Putte and Straßer, 2012]. Hence, just as for minimal abnormality, we can select the $\sqsubset$-minimal models of a premise set and define semantic consequence in terms of those models. Then it is again a matter of routine to show that this consequence relation satisfies all the nice properties of the standard format.

For an illustration of this format of ALs, let us suppose that *prima facie* obligations come in various degrees $i \in \mathbb{N}$ of importance, where degree 1 is most important, degree 2 is slightly less important, etc. Let $\mathsf{O}^{\mathsf{p}}_i A$ denote that A is *prima facie* obligatory, with degree i. Then intuitively, we expect that from $\{\mathsf{O}^{\mathsf{p}}_1 p, \mathsf{O}^{\mathsf{p}}_2 q, \mathsf{O}^{\mathsf{p}}_2 r, \neg\mathsf{O}(p \wedge q)\}$ we can derive $\mathsf{O}p$ but not $\mathsf{O}q$. Moreover, we also expect $\mathsf{O}r$ to be derivable, since r is not involved in the conflict. This is exactly the result we obtain if we define our sequence of sets of abnormalities as $\langle\{\mathsf{O}^{\mathsf{p}}_i A \wedge \neg\mathsf{O}A\}\rangle_{i \in \mathbb{N}}$.

The format of lexicographic ALs is relatively new; the first ideas for it date back to 2010. It has been applied to deontic logic in [Van De Putte and Straßer, 2013], where a lexicographic variant of the logic from [Meheus *et al.*, 2012] is proposed.

3.5 Further reading

The first ALs were developed a little before 1980 by Diderik Batens, as a new, "dialectical" aproach to (non-explosive) reasoning with inconsistent theories.[24] Nowadays these logics are called "inconsistency-adaptive logics" – more on them in Section 7.[25]

From its first days, this research was pluralist in the sense that various (monotonic) paraconsistent logics were used to define ALs. Around the mid 1990s, the idea emerged that besides inconsistency, various other

[24] In [Batens, 1986], Batens refers to an (unpublished) manuscript from 1979, "Dynamische processen en dialectische logica's", as the first paper on this subject.

[25] The term "adaptive" appears to be introduced in 1981 [Batens, 1986].

types of "abnormality" with respect to classical (propositional or first order) logic can be used as a basis to define ALs – see e.g. [Batens, 1997]. The resulting logics are nowadays called "corrective ALs", in contradistinction to "ampliative" ALs, which only saw light around 2000.[26] The latter are, roughly, ALs that characterize a given type of inference which goes *beyond* one's chosen standard of deduction (usually first order **CL**), such as compatibility [Batens and Meheus, 2000], inductive generalization [Batens, 2011], abduction [Meheus *et al.*, 2002; Beirlaen and Aliseda, 2014], etc.

The notion of an adaptive strategy was only fully developed in the 1990s – see in particular [Batens, 1999a]. Before that, only the proof theory of *reliability* and the semantics of *minimal abnormality* were known.

The standard format as presented in this section, was introduced in [Batens, 2007]. Its further development in turn facilitated applications in various new areas during the last decade, ranging from foundations of set theory [Verdée, 2013], over causal discovery [Leuridan, 2009; Beirlaen *et al.*, 2018], to deontic logic.

For a recent and compact introduction into ALs (with a focus on their application to paraconsistent reasoning), we refer to [Batens, 2015]. A thorough discussion of the standard format and several of its generalizations can be found in Part I of [Straßer, 2014]. Slightly older papers that present the basics of ALs are [Batens, 2001] and [Batens, 2007].

ALs have been compared to various other generic frameworks for defeasible and/or non-monotonic reasoning in the past, including Makinson's *default assumption consequence relations* [Van De Putte, 2013], abstract argumentation [Straßer and Šešelja, 2010], and modal logics [Allo, 2013]. There is also an interesting line of research on the relation between ALs and Rescher-Manor consequence relations for "contextualized" reasoning with inconsistent premises [Rescher and Manor, 1970]. In fact, the logics $\mathbf{SDL^r_p}$, $\mathbf{SDL^m_p}$, and $\mathbf{SDL^n_p}$ can be seen as adaptive variants of the *Free*, the *Strong*, and the *Weak* Rescher-Manor consequence relation respectively [Meheus *et al.*, 2016].

4 Revisionist adaptive deontic logics

The logics $\mathbf{SDL^r_p}$ and $\mathbf{SDL^m_p}$ from Section 3 reserve the **SDL**-operator O for actual obligations, while they allow for the non-trivial formalization of conflicting (*prima facie*) obligations in terms of the new operator $\mathsf{O^p}$. Via this grammatical enrichment, we obtain a conflict-tolerant adaptive

[26] See e.g. [Meheus *et al.*, 2002] for a discussion of this distinction.

logic, without having to revise any of the core principles of **SDL**. Indeed, $\mathbf{SDL}^{\mathbf{x}}_{\mathbf{p}}$ is built on top of $\mathbf{SDL}_{\mathbf{p}}$, which is in turn an extension of **SDL**.

Instead of extending the grammar of **SDL** while keeping its core principles intact, we may also accommodate conflicts by keeping the grammar of **SDL** intact while giving up some of its core principles. This means that we *revise* the underlying logic, to use the terminology from [Goble, 2013]. We therefore call the adaptive logics based on such "weak" deontic logics *revisionist adaptive deontic logics*. The aim of sections 5–7 is to present and discuss this branch of ALs.

We provide some general insight into the various types of revisionist (adaptive) deontic logics that are on the market in Section 4.1. Next, we will introduce some conceptual machinery that allows us to compare and evaluate such logics (Section 4.2).

4.1 SDL: three ways of giving it up (while keeping it)

If we are to reason non-trivially in the face of conflicting obligations, we need to give up at least some part of **SDL**. For the time being, let us focus on conflicts of the type $\mathsf{O}A \wedge \mathsf{O}\neg A$ (we will consider several other types below). First, if the logic of $\neg$ is classical, then the (D)-axiom needs to be given up in order to avoid that everything follows from $\mathsf{O}A \wedge \mathsf{O}\neg A$. This means we are left with the minimal normal modal logic **K**, which is fully characterized by **CL**, the rule of necessitation (N) and the normality schema (K).

But giving up (D) alone will not do. As soon as (Agg), (Inh), and *Ex Contradictione Quodlibet* (ECQ) are valid, deontic conflicts result in deontic explosion, i.e. the conclusion that everything is obligatory:[27]

$$\mathsf{O}A, \mathsf{O}\neg A \vdash \mathsf{O}B \qquad \text{(DEX)}$$

Suppose $\mathsf{O}A$ and $\mathsf{O}\neg A$. By (Agg), $\mathsf{O}(A \wedge \neg A)$. By (ECQ) and (Inh), $\mathsf{O}B$. Since all three of these principles are derivable within **K**, deontic conflicts imply deontic explosion also in this minimal logic.

So at least one of (Agg), (Inh), or (ECQ) has to go. It can be shown – and will be shown in the next three sections – that giving up either (Agg), or (Inh), or (ECQ) is sufficient in order to accommodate conflicts of the type $\mathsf{O}A \wedge \mathsf{O}\neg A$.[28] So in the remainder we will focus on these

[27] (ECQ) is the (classically valid) inference from $A, \neg A$ to arbitrary B.

[28] One may of course give up even more principles, but we will focus on the simple cases where only one of the three is given up. All that we write on revisionist deontic logics and their adaptive extensions applies *mutatis mutandis* to such weaker logics.

three principles, rather than on the "official" characterization of **SDL** in terms of (N), (K) and (D).

In Section 5, we will consider deontic logics that are obtained by giving up (Inh).[29] This means that e.g. $\mathsf{O}(A \wedge B)$ does not imply $\mathsf{O}A$, and $\mathsf{O}A$ does not imply $\mathsf{O}(A \vee C)$ in these logics, absent further information about A, B, and C. As a result, $\mathsf{O}(A \wedge B)$ can be true for conflicting (i.e., mutually incompatible) A and B, but this need not imply that $\mathsf{O}C$ is true for any arbitrary (non-contradictory) C.

Section 6 is concerned with conflict-tolerant deontic logics that invalidate (Agg). Thus, in these logics, $\mathsf{O}A$ and $\mathsf{O}B$ can be true without $\mathsf{O}(A \wedge B)$ being true. As a result, the step from $\mathsf{O}A \wedge \mathsf{O}\neg A$ to $\mathsf{O}(A \wedge \neg A)$ is blocked and we cannot get to the conclusion that any B is obligatory.

Finally, Section 7 focuses on alternative, weaker accounts of negation, which invalidate (ECQ). This allows us to keep (D).

So there are several, well-studied ways to avoid (DEX) and thus to accommodate deontic conflicts within a formal logic. However, giving up principles of **SDL** comes at a price. As we will show below, these principles are at the heart of intuitively plausible patterns of inference – see Section 4.2 for a number of examples. Giving up the principles means that one either has to deny head-on the validity of those inferences, or to explain them as enthymatic arguments, i.e. arguments with a number of tacit, hidden premises. Even if such a strategy is successful to some extent, it turns out very difficult to develop a general logical (and philosophically justifiable) procedure that allows one to obtain such tacit premises for a given case.

Going adaptive allows us to give up principles, whilst keeping them *as much as possible*, i.e., as long as they do not lead to deontic explosion. The core idea behind revisionist adaptive deontic logics is to start from a monotonic, conflict-tolerant deontic logic **L** and to try to apply the missing **SDL**-rule(s) in a premise-sensitive, defeasible way, thus steering a middle course between the excesses of **SDL** and the inferential weakness of **L**.

Before we continue, an important side-remark is in place. In [Goble, 2013, Sect. 5.4], Goble also develops two new, monotonic conflict-tolerant deontic logics that are inferentially very powerful, in the sense that they validate (a variant of) (Agg), (DDS), and (Dist). The basic idea behind these logics is to give up the principle of extensionality (RE), and to opt for a weaker notion of "analytic equivalence" instead.

[29] In some but not all of these logics, also (Agg) is restricted. In all of them, classical logic is preserved for the connectives and replacement of equivalents (RE) holds.

In recent (unpublished) work, Anglberger and Korbmacher have developed a semantics for the resulting logics, based on truthmaker semantics for hyperintensional logics [Fine, 2016]. We will not discuss these new systems in the present chapter, since it is as yet unclear whether and how sensible adaptive logics based on them could be developed.

4.2 Criteria for comparison and evaluation

When discussing and comparing the ALs defined in the next three sections, we will look at two aspects in particular. First, we will consider various types of deontic conflicts, and compare the logics in terms of which of these types they can accommodate properly. Second, we look at how the logics behave with respect to specific benchmark examples known from the literature.

Explosion principles In the specific context of conflict-tolerant deontic logics, it is common to demand some additional consistency constraints on top of the consistency preservation property from Theorem 3.11. In particular, we want to take great care to avoid the validity of *explosion principles*, i.e. principles according to which a set of arbitrary formulas is derivable given a (specific type of) normative conflict. These can come in various types, as we now explain.

We already referred to the principle of deontic explosion (DEX) in Section 4.1. In [Straßer and Beirlaen, 2011], some more refined explosion principles are specified that serve as touchstones for measuring the conflict-tolerance of various deontic logics. Here are some examples:[30]

$$\mathsf{O}A, \mathsf{O}\neg A \vdash \mathsf{O}B \vee \mathsf{O}\neg B \tag{2}$$

$$\mathsf{O}A, \mathsf{O}\neg A \vdash \mathsf{O}B \vee \mathsf{P}B \tag{3}$$

$$\mathsf{O}A, \mathsf{O}\neg A \vdash \mathsf{O}B \vee \neg\mathsf{O}\neg B \tag{4}$$

$$\mathsf{O}A, \mathsf{O}\neg A \vdash \mathsf{P}B \tag{5}$$

Principles (2)-(5) weaken the right-hand side of (DEX). We can devise further – equally undesirable – explosion principles by strengthening its left-hand side via the addition of logically unrelated information. For instance, where γ is any subset of $\{\mathsf{O}D, \neg\mathsf{O}\neg D, \mathsf{P}E, \neg\mathsf{O}\neg E, \neg\mathsf{O}F, \neg\mathsf{O}\neg F, \mathsf{P}G, \mathsf{P}\neg G\}$,

$$\{\mathsf{O}A, \mathsf{O}\neg A\} \cup \gamma \vdash \mathsf{O}B \vee \mathsf{P}B \tag{6}$$

[30] Recall that we treat O and P as primitive operators unless stated otherwise; cf. Section 2.

More fine-grained explosion principles may be obtained by stipulating that principles like (2)-(6) are avoided even for B that satisfy certain additional constraints. For instance, Goble showed that the following principle is valid in deontic logics which restrict (Agg) to conjunctions of jointly consistent obligations [Goble, 2005]:

$$\text{If } \nvdash \neg B, \text{ then } \mathsf{O}A, \mathsf{O}\neg A \vdash \mathsf{O}B \tag{7}$$

The above forms of explosion are all still limited in (at least) one sense, in that they are focused on binary conflicts between obligations, i.e. formulas of the form $\mathsf{O}A \wedge \mathsf{O}\neg A$. There seems to be no reason to us as to why one should focus solely on such types of conflicts between norms, ignoring all others. For instance, there seems to be no logical reason why self-contradictory norms should be excluded – if an authority can issue mutually incompatible commands, then why can't it issue (highly complex but) self-contradictory commands as well? Likewise, why not consider conflicts between obligations and permissions?

Consider the following variant of an example from [Hansen, 2014, p. 305]: a couple you know is having a party. One of them leaves a message: "I am sorry, you cannot come – it's close friends only." The other also leaves a message: "you can surely come to the party if you like – there will anyway be plenty of food for everyone." Absent further information, the resulting norms can best be formalized as $\neg\mathsf{P}p$ and $\mathsf{P}p$, where p stands for "go to the party". Even if we assume that O and P are interdefinable, this does not result in a conflict of the form $\mathsf{O}A \wedge \mathsf{O}\neg A$, but rather in a direct contradiction, i.e. $\mathsf{O}A \wedge \neg\mathsf{O}A$.

So all in all, there seem to be reasons for taking into account explosion principles such as the following:

$$\mathsf{O}A, \mathsf{P}\neg A \vdash \mathsf{O}B \tag{8}$$

$$\mathsf{O}(A \wedge \neg A) \vdash \mathsf{O}B \tag{9}$$

Candidate conflict-tolerant deontic logics should be tested not only for the validity of (DEX), but also for the validity of more refined principles like (2)-(7) above. In doing so, we do not consider it the task of any such logic to invalidate all forms of explosion; rather, we treat the explosion principles as a useful way to compare and classify given deontic logics.

In the next two sections, we will focus on the following explosion principles – apart from (DEX):

$$\mathsf{O}(A \wedge \neg A) \vdash \mathsf{O}B \qquad \text{(DEX-O}\bot\text{)}$$
$$\mathsf{P}(A \wedge \neg A) \vdash \mathsf{P}B \qquad \text{(DEX-P}\bot\text{)}$$
$$\mathsf{O}A \wedge \mathsf{P}\neg A \vdash B \qquad \text{(DEX-OP}\neg\text{)}$$
$$\mathsf{O}A \wedge \neg\mathsf{P}A \vdash B \qquad \text{(DEX-O}\neg\mathsf{P}\text{)}$$

We choose these five principles since they allow us to compare the (non)explosive behavior of the various logics discussed below in a succinct way. In Section 7 we will consider some additional forms of explosion that can be avoided by using paraconsistent deontic logics.

Benchmark examples. Research in the fields of deontic logic and non-monotonic logic is to a large extent driven by a relatively small set of benchmark examples aimed at testing the formal system in question (the reader may be familiar with Tweety the penguin, the good Samaritan, and the gentle murderer, just to name a few). When faced with such examples, counter-intuitive outcomes are taken to reflect badly on a formal system, so these benchmark examples provide a criterion for checking whether a formal system meets our informal intuitions.

A warning is in order here, however. The fact that a formal system provides intuitive outcomes for the relevant benchmark examples is not a sufficient condition for positively evaluating the system in question. For instance, the system may be devised in an *ad hoc* manner to deal specifically with a small set of examples, at the cost of violating one or more rationality postulates. Moreover, some of these examples may reflect intuitions on which not everyone agrees, leaving room for dispute. In some cases the fact that our logic does *not* give us the expected outcome for some concrete example may inform us that our intuitions are perhaps incoherent, whence this is not in itself a sufficient reason to reject the logic. So, as was the case with explosion principles, we will use our benchmark examples as means to classify given logics, not as absolute criteria for their usefulness.[31]

With this warning in mind, let us list a number of examples which have been used to evaluate conflict-tolerant deontic logics studied in the literature. For each of them, we indicate some of the basic **SDL**-principles which allow us to infer the conclusion from the given premises. We use (CL) as a generic name for all inferences that are **CL**-valid.

[31] For a critical discussion of the use of examples as intuition-pumps in the evaluation of logics for defeasible reasoning, see [Prakken, 2002].

1. *The Smith Argument.* — (Agg), (Inh), (CL)

   (i) Smith ought to fight in the army or perform alternative service to his country ($\mathsf{O}(f \vee s)$).
   (ii) Smith ought not to fight in the army ($\mathsf{O}\neg f$).
   ∴ (iii) Smith ought to perform alternative service to his country ($\mathsf{O}s$).

2. *The Jones Argument.* — (Inh), (CL)

   (i) Jones ought to tell a joke and sing a song ($\mathsf{O}(j \wedge s)$).
   ∴ (ii) Jones ought to tell a joke ($\mathsf{O}j$).

3. *The Roberts Argument, version 1.* — (Inh), (CL)

   (i) Roberts ought to pay federal taxes and register for national service ($\mathsf{O}(t \wedge r)$).
   (ii) Roberts ought not to pay federal taxes but volunteer to help the homeless in his community ($\mathsf{O}(\neg t \wedge v)$).
   ∴ (iii) Roberts ought to register for national service and ought to volunteer to help the homeless ($\mathsf{O}r \wedge \mathsf{O}v$).

4. *The Roberts Argument, version 2.* — (Inh), (CL), (Agg)

   (i) Roberts ought to pay federal taxes and register for national service ($\mathsf{O}(t \wedge r)$).
   (ii) Roberts ought not to pay federal taxes but volunteer to help the homeless in his community ($\mathsf{O}(\neg t \wedge v)$).
   ∴ (iii) Roberts ought to register for national service and volunteer to help the homeless ($\mathsf{O}(r \wedge v)$).

5. *The Thomas Argument.* — (Inh), (Agg), (CL)

   (i) Thomas ought to pay federal taxes and either fight in the army or perform alternative service to his country ($\mathsf{O}(t \wedge (f \vee s))$).
   (ii) Thomas ought neither to pay federal taxes nor fight in the army ($\mathsf{O}(\neg t \wedge \neg f)$).
   ∴ (iii) Thomas ought to perform alternative service to his country ($\mathsf{O}s$).

6. *The Natascha Argument, version 1.* — (K) / (Inh), (Agg), (CL)

   (i) Natascha ought to take Sarah to the concert ($\mathsf{O}s$).
   (ii) Natascha ought to take Martin to the concert ($\mathsf{O}m$).
   (iii) It is not the case that Natascha ought to take Sarah *and* Martin to the concert ($\neg\mathsf{O}(s \wedge m)$).
   (iv) If she takes Sarah, she ought to buy an extra ticket ($\mathsf{O}(s \supset t)$).
   (v) If she takes Martin, she ought to buy an extra ticket ($\mathsf{O}(m \supset t)$).
   ∴ (vi) Natascha ought to buy an extra ticket ($\mathsf{O}t$).

7. *The Natascha Argument, version 2.* — (K) / (Inh), (Agg), (CL)

   (i) Natascha ought to take Sarah to the concert ($\mathsf{O}s$).
   (ii) Natascha ought to take Martin to the concert ($\mathsf{O}m$).
   (iii) Natascha ought not to take Sarah *and* Martin to the concert ($\mathsf{O}\neg(s \wedge m)$).
   (iv) If she takes Sarah, she ought to buy an extra ticket ($\mathsf{O}(s \supset t)$).
   (v) If she takes Martin, she ought to buy an extra ticket ($\mathsf{O}(m \supset t)$).
   ∴ (vi) Natascha ought to buy an extra ticket ($\mathsf{O}t$).

The Smith argument was first presented by Horty [1994; 1997; 2003; 2012]; the name 'Smith' is due to Goble [2014; 2013]. The Jones, Roberts, and Thomas arguments are variations on examples from [Goble, 2014; Goble, 2013]. The Natascha argument is new.

The validity of these arguments is not undisputed. The Jones argument, for instance, which concerns the application of the inheritance principle (Inh), has been called into question [Goble, 1990a; Hansen, 2013; Parent and van der Torre, 2014]. The Natascha argument concerns the derivation of a so-called *floating conclusion*, a conclusion entailed by each of two mutually conflicting obligations. The status of such conclusions is debatable.[32]

[32]See [Horty, 2002; Makinson and Schlechta, 1991; Prakken, 2002] for arguments pro and contra the derivation of floating conclusions in non-monotonic logic. In a moral context, the derivability of floating conclusions has been defended by Brink [1994].

In both versions of the Natascha argument, the idea behind the third premise is that for some reason or another, Natascha cannot possibly take both Sarah and Martin to the concert — e.g. because there is only one additional ticket left at the counter. In the absence of alethic modalities, we translate information concerning what is (im)possible directly into the language of **SDL**. While the first version of this argument relies on the principle of "ought implies can" (OIC) and contraposition, the second relies on the stronger principle of "permitted implies can" (PIC), interdefinability of O and P, and contraposition. Both (OIC) and (PIC) are controversial.[33] However, here we focus merely on the formal premises as such, not on the question whether they represent the example in the most natural way.

5 Adaptive inheritance

The first type of conflict-tolerant deontic logics mentioned in Section 4.1 is obtained by giving up or weakening the rule of inheritance (Inh). In the present section, we discuss one specific subclass of such logics, showing how they can be strengthened by going adaptive.

5.1 Logics with unconflicted inheritance

Restricting inheritance In a number of papers, Goble presented the **LUM**-family of deontic logics.[34] The language of these logics is just that of **SDL**, with P defined as the dual of O. The logics in the **LUM**-family do not simply reject inheritance, but replace it with a weaker principle that accounts for a number of intuitive applications of (Inh). This requires some explanation.

Let $\mathsf{U}A =_{\mathsf{df}} \neg(\mathsf{O}A \wedge \mathsf{O}\neg A)$ denote that A is unconflicted. All **LUM**-systems extend **CL** with the necessitation rule (N), the replacement of equivalents rule (RE), as well as the following rule of "unconflicted" inheritance (RUM):

$$\text{If } A \vdash B, \text{ then } \mathsf{U}A, \mathsf{O}A \vdash \mathsf{O}B \qquad \text{(RUM)}$$

(RUM) allows for those applications of the inheritance rule (Inh) which involve only unconflicted obligations. In terms of permission, the

[33] See [Vranas, 2007] for a comprehensive discussion of the first of these two principles.

[34] We adopt the presentation and nomenclature from [Goble, 2014]. For more details and references, we refer to Section 5.3.

rule states that whenever A is both obligatory and permitted, then whatever is logically weaker than A is also obligatory. This rule is therefore also sometimes referred to as "permitted inheritance" (RPM).

In addition to (N), (RE), and (RUM), the systems in the **LUM**-family are defined in terms of (a selection among) (P), (Agg), and "consistent" and "permitted" aggregation rules (C-Agg) and (P-Agg):

$$\text{If } \nvdash \neg(A \wedge B) \text{ then } \mathsf{O}A, \mathsf{O}B \vdash \mathsf{O}(A \wedge B) \qquad \text{(C-Agg)}$$

$$\mathsf{P}A, \mathsf{P}B, \mathsf{O}A, \mathsf{O}B \vdash \mathsf{O}(A \wedge B) \qquad \text{(P-Agg)}$$

Note that, since P is the dual of O, the antecedent of (P-Agg) just means that A and B are obligatory, and that neither of their negations are obligatory. The systems **LUM.a**-**LUM.c** extend **CL** by adding:

LUM.a: (N), (RE), (RUM), (Agg)
LUM.b: (N), (RE), (RUM), (P), (C-Agg)
LUM.c: (N), (RE), (RUM), (P), (P-Agg)

A semantics for these three logics can easily be obtained, following the well-known generalization of Kripke-semantics into neighbourhood semantics – cf. [Chellas, 1980, Chapters 7 & 8] and [Segerberg, 1971]. Say a **LUM**-model is of the type $M = \langle W, w_0, n_{\mathsf{O}}, v\rangle$, where W is a non-empty set of worlds, $w_0 \in W$ is the actual world, $n_{\mathsf{O}} : W \to \wp(\wp(W))$ maps each world $w \in W$ to the set of *obligatory propositions at* w, and v is a valuation function. The semantic clause for O in such models reads:

(SC-O) $M, w \models \mathsf{O}A$ iff $|A|_M \in n_{\mathsf{O}}(w)$

Truth in a model is defined as usual, viz. as truth at w_0; semantic consequence is defined by quantifying over all models in which the premises are true.

This gives us the minimal classical modal logic **E**, which is characterized fully by adding (RE) to **CL**. Imposing a number of restrictions on such models, we obtain the additional axioms and rules listed above. These conditions are:

(CO-RUM) if $X \in n_{\mathsf{O}}(w)$, $W \setminus X \notin n_{\mathsf{O}}(w)$, and $X \subseteq Y$, then $Y \in n_{\mathsf{O}}(w)$
(CO-N) $W \in n_{\mathsf{O}}(w)$
(CO-P) $\emptyset \notin n_{\mathsf{O}}(w)$
(CO-Agg) if $X \in n_{\mathsf{O}}(w)$ and $Y \in n_{\mathsf{O}}(w)$, then $X \cap Y \in n_{\mathsf{O}}(w)$
(CO-C-Agg) if $X \in n_{\mathsf{O}}(w)$, $Y \in n_{\mathsf{O}}(w)$, and $X \cap Y \neq \emptyset$, then $X \cap Y \in n_{\mathsf{O}}(w)$

(CO-P-Agg) if $X \in n_{\mathsf{O}}(w)$, $Y \in n_{\mathsf{O}}(w)$, $W \setminus X \notin n_{\mathsf{O}}(w)$, and $W \setminus Y \notin n_{\mathsf{O}}(w)$, then $X \cap Y \in n_{\mathsf{O}}(w)$

For an extensive comparison and discussion of the various **LUM**-logics, we refer to [Goble, 2013, Sect. 5.3]. In the remainder, we will focus on ALs obtained from them.

Going adaptive To understand the specific motivation for going adaptive in the case of the **LUM**-logics, it will be useful to reconsider the benchmark examples from Section 4.2. The Smith and Jones arguments are invalid in all three of the **LUM**-logics, but valid once we add the premises $\mathsf{U}\neg f$ (for the Smith argument) and $\mathsf{U}(j \wedge s)$ (for the Jones argument). The Roberts and Thomas arguments are more problematic. In the Roberts argument, for instance, we cannot just add the premises $\mathsf{U}(t \wedge r)$ and $\mathsf{U}(\neg t \wedge v)$ in order to render the argument valid, since doing so would trivialize the premise set.[35]

More generally, it is problematic that in the **LUM**-systems we need to add the 'tacit' information that a formula is unconflicted before we can apply the restricted distribution rule. This worry was first raised in [Straßer *et al.*, 2012], and acknowledged by Goble:

> For one thing, the additional non-conflict condition on the distribution rule seems rather *ad hoc*; there is little to recommend it except its success in disarming deontic explosion. For another, it seems risky to try to account for the plausibility of arguments by considering them enthymematic for straight-forwardly valid arguments. In context it may be all right to accept the alleged tacit premise, but we cannot rely on that. With more complicated arguments it might be quite uncertain what unspoken premises of non-conflict are implicitly present [Goble, 2014, pp. 210-211].

Both problems can be overcome by strengthening the **LUM**-systems within the adaptive logics framework. On the one hand, we can validate all those applications of distribution that do not lead to deontic explosion. On the other hand, it is the logic itself that fixes which applications of distribution are tolerable; no interference of any user is

[35] For Roberts, first note that $\vdash (t \wedge r) \supset \neg(\neg t \wedge v)$. By (RPM), $\neg\mathsf{O}\neg(t \wedge r) \supset (\mathsf{O}(t \wedge r) \supset \mathsf{O}\neg(\neg t \wedge v))$ or, equivalently, $\mathsf{O}\neg(t \wedge r) \vee (\mathsf{O}(t \wedge r) \supset \mathsf{O}\neg(\neg t \wedge v))$. By premises (i) and (ii) of the Roberts argument, we get $(\mathsf{O}(t \wedge r) \wedge \mathsf{O}\neg(t \wedge r)) \vee (\mathsf{O}(\neg t \wedge v) \wedge \mathsf{O}\neg(\neg t \wedge v))$ by **CL**. So adding $\mathsf{U}(t \wedge r)$ and $\mathsf{U}(\neg t \wedge v)$ would make the argument **CL**-inconsistent. For Thomas the argument is analogous.

required for this. We explain how this works below, focusing on the adaptive extensions of the logic **LUM.a**. For the other logics in this family, the difficulties and properties are roughly analogous. We will point out salient differences as we go along.

The logics LUM.a$^{\mathbf{x}}$ A natural way of strengthening Goble's **LUM**-systems is to work under the assumption that obligations are unconflicted, so that an obligation $\mathsf{O}A$ behaves abnormally in case it is conflicted, i.e. in case $\neg\mathsf{U}A$ or, equivalently, $\mathsf{O}A \wedge \mathsf{O}\neg A$:

$$\Omega = \{\mathsf{O}A \wedge \mathsf{O}\neg A \mid A \in \mathcal{W}\}$$

The logic **ADPM.1$^{\mathbf{r}}$** from [Straßer *et al.*, 2012] is the AL defined by the triple $\langle$**LUM.a**$, \Omega,$ reliability$\rangle$. In an **ADPM.1$^{\mathbf{r}}$**-proof, (Inh) can be applied via the conditional rule RC, assuming that the obligations involved are not conflicted:

1	$\mathsf{O}(p \wedge q)$	Prem	$\emptyset$
2	$\mathsf{O}p$	1; RC	$\{\mathsf{O}(p \wedge q) \wedge \mathsf{O}\neg(p \wedge q)\}$

The conditional derivation at line 2 is legitimate in view of the **LUM.a**-valid inference

$$\mathsf{O}(p \wedge q) \vdash \mathsf{O}p \vee (\mathsf{O}(p \wedge q) \wedge \mathsf{O}\neg(p \wedge q)) \tag{10}$$

Unfortunately, Goble pointed out that **ADPM.1$^{\mathbf{r}}$** suffers from a problem [Goble, 2014, Sect. 4.3.1]. Although we can indeed apply distribution conditionally in **ADPM.1$^{\mathbf{r}}$**, the corresponding application of RC in the proof is marked as soon as a (possibly unrelated) conflict follows from the premise set. The problem is best illustrated by means of a simple example.

1	$\mathsf{O}(p \wedge q)$	Prem	$\emptyset$
2	$\mathsf{O}r$	Prem	$\emptyset$
3	$\mathsf{O}\neg r$	Prem	$\emptyset$
4	$\mathsf{O}p$	1; RC	$\{\mathsf{O}(p \wedge q) \wedge \mathsf{O}\neg(p \wedge q)\}\checkmark$
5	$(\mathsf{O}(p \wedge q) \wedge \mathsf{O}\neg(p \wedge q)) \vee (\mathsf{O}(p \wedge r) \wedge \mathsf{O}\neg(p \wedge r)) \vee (\mathsf{O}(p \wedge \neg r) \wedge \mathsf{O}\neg(p \wedge \neg r))$	1-3; RU	$\emptyset$

The Dab-formula derived at line 5 is minimal at this stage of the

proof, and causes the marking of line 4.[36] This Dab-formula is a minimal Dab-consequence of the premise set $\{\mathsf{O}(p \wedge q), \mathsf{O}r, \mathsf{O}\neg r\}$. Consequently, there is no extension of this proof in which line 4 is unmarked, and hence

$$\mathsf{O}(p \wedge q), \mathsf{O}r, \mathsf{O}\neg r \nvdash_{\mathbf{ADPM.1^r}} \mathsf{O}p \tag{11}$$

The same holds if we use the minimal abnormality strategy instead of reliability (the reasoning is analogous):

$$\mathsf{O}(p \wedge q), \mathsf{O}r, \mathsf{O}\neg r \nvdash_{\mathbf{ADPM.1^m}} \mathsf{O}p \tag{12}$$

This problem generalizes: in the presence of a conflict between two obligations, we can construct minimal Dab-formulas containing abnormalities pertaining to seemingly unrelated and unproblematic formulas, blocking unproblematic applications of RC. The logics $\mathbf{ADPM.1^r}$ and $\mathbf{ADPM.1^m}$ are therefore called *flip-flops* [Batens, 2007]. In the absence of conflicts, their consequence set is the same as their ULL, namely **SDL**.[37] As soon as one conflict is present, however, their consequence set collapses into that of their lower limit logic **LUM.a**.

There is a natural fix to this flip-flop problem, due to Goble [Goble, 2014]. Let $\mathrm{S}(A)$ denote the set of all subformulas of A (including A itself). Where $\mathrm{S}(A) = \{B_1, \ldots, B_n\}$, we define[38]

$$\sharp(A) = (\mathsf{O}B_1 \wedge \mathsf{O}\neg B_1) \vee \ldots \vee (\mathsf{O}B_n \wedge \mathsf{O}\neg B_n)$$

Following Goble, we let $\mathbf{LUM.a^r} = \langle \mathbf{LUM.a}, \Omega^{\mathrm{S}}, \text{reliability} \rangle$, where[39]

$$\Omega^{\mathrm{S}} = \{\sharp(A) \mid A \in \mathcal{W}\}$$

In an $\mathbf{LUM.a^r}$-proof, the formula derived at line 5 of our proof above is no longer a Dab-formula. Rather, we obtain the following proof:

1	$\mathsf{O}(p \wedge q)$	Prem	$\emptyset$
2	$\mathsf{O}r$	Prem	$\emptyset$
3	$\mathsf{O}\neg r$	Prem	$\emptyset$
4	$\mathsf{O}p$	1; RC	$\{\sharp(p \wedge q)\}$
5	$\sharp(p \wedge q) \vee \sharp(p \wedge r) \vee \sharp(p \wedge \neg r)$	1-3; RU	$\emptyset$
6	$\sharp(p \wedge r)$	2,3; RU	$\emptyset$
7	$\sharp(p \wedge \neg r)$	2,3; RU	$\emptyset$

[36] By (10), $\mathsf{O}p \vee (\mathsf{O}(p \wedge q \wedge \mathsf{O}\neg(p \wedge q))$. Suppose $\mathsf{O}p$. Then (i) by (Agg), $\mathsf{O}(p \wedge r)$ and, by (RUM) and **CL**, $\mathsf{O}\neg(p \wedge \neg r) \vee (\mathsf{O}(p \wedge r) \wedge \mathsf{O}\neg(p \wedge r))$; analogously (ii) by (Agg), $\mathsf{O}(p \wedge \neg r)$ and, by (RUM) and **CL**, $\mathsf{O}\neg(p \wedge r) \vee (\mathsf{O}(p \wedge \neg r) \wedge \mathsf{O}\neg(p \wedge \neg r))$. Altogether, by **CL**, $(\mathsf{O}(p \wedge q) \wedge \mathsf{O}\neg(p \wedge q)) \vee (\mathsf{O}(p \wedge r) \wedge \mathsf{O}\neg(p \wedge r)) \vee (\mathsf{O}(p \wedge \neg r) \wedge \mathsf{O}\neg(p \wedge \neg r))$.

[37] It was shown in [Straßer *et al.*, 2012, Th. 7] that **SDL** is the ULL of $\mathbf{ADPM.1^r}$.

[38] Our expression $\sharp(A)$ is equivalent to the negation of Goble's expression $\mho(A)$ in [Goble, 2014; Goble, 2013]. Note that $\sharp$ is not a (modal or other) operator but just a symbol that allows us to abbreviate a formula.

[39] Goble uses the name $\mathbf{ALUM^r}$ for the logic that we call $\mathbf{LUM.a^r}$.

The abnormalities $\sharp(p \wedge q)$, $\sharp(p \wedge r)$, and $\sharp(p \wedge \neg r)$ denote the formulas (13), (14), and (15) respectively:

$$(\mathsf{O}(p \wedge q) \wedge \mathsf{O}\neg(p \wedge q)) \vee (\mathsf{O}p \wedge \mathsf{O}\neg p) \vee (\mathsf{O}q \wedge \mathsf{O}\neg q) \tag{13}$$

$$(\mathsf{O}(p \wedge r) \wedge \mathsf{O}\neg(p \wedge r)) \vee (\mathsf{O}p \wedge \mathsf{O}\neg p) \vee (\mathsf{O}r \wedge \mathsf{O}\neg r) \tag{14}$$

$$(\mathsf{O}(p \wedge \neg r) \wedge \mathsf{O}\neg(p \wedge \neg r)) \vee (\mathsf{O}p \wedge \mathsf{O}\neg p) \vee (\mathsf{O}\neg r \wedge \mathsf{O}\neg\neg r) \vee (\mathsf{O}r \wedge \mathsf{O}\neg r) \tag{15}$$

The inference made at line 4 is legitimate in view of the **LUM.a**-valid inference

$$\mathsf{O}(p \wedge q) \vdash \mathsf{O}p \vee \sharp(p \wedge q) \tag{16}$$

Since $\sharp(p \wedge r)$ and $\sharp(p \wedge \neg r)$ are **LUM.a**-derivable from the premises $\mathsf{O}r$ and $\mathsf{O}\neg r$, the Dab-formula derived at line 5 of the proof is not minimal at stage 7. Consequently, line 4 is unmarked at this stage. As opposed to **ADPM.1$^{\mathbf{r}}$** and **ADPM.1$^{\mathbf{m}}$**, the logics **LUM.a$^{\mathbf{r}}$** and **LUM.a$^{\mathbf{m}}$** lead to the following desirable outcome:

$$\mathsf{O}(p \wedge q), \mathsf{O}r, \mathsf{O}\neg r \vdash_{\mathbf{LUM.a^r}} \mathsf{O}p \tag{17}$$

$$\mathsf{O}(p \wedge q), \mathsf{O}r, \mathsf{O}\neg r \vdash_{\mathbf{LUM.a^m}} \mathsf{O}p \tag{18}$$

5.2 Evaluating the logics

Explosion principles The adaptive logics based on the **LUM**-family are conflict-tolerant to the same extent as their respective lower limit logics. This means, for a start, that (DEX) is invalid in all of them. Since they are **CL**-based and in view of the interdefinability of O and P, they also accommodate conflicts of the form $\mathsf{O}A \wedge \neg\mathsf{P}A$, which simply reduce to conflicts between obligations.

However, the logics do not tolerate the other types of deontic conflicts that were discussed in Section 4.2. While $\mathsf{O}(A \wedge \neg A)$ is consistent in **LUM.a** – and hence also in **LUM.a$^{\mathbf{x}}$**, it is inconsistent in each of **LUM.b** and **LUM.c** in view of the (P)-axiom. It follows that ALs based on the latter two logics cannot make sense of self-contradictory obligations. Also, all the (adaptive) **LUM**-logics trivialize conflicts of the form $\mathsf{O}A \wedge \mathsf{P}\neg A$, as these reduce to plain contradictions in view of ($\mathsf{Def_P}$) and (RE). Finally, $\mathsf{P}(A \wedge \neg A)$ (which is equivalent to $\neg\mathsf{O}(\neg A \vee A)$) is also trivial in these logics, in view of the necessitation rule (N).

Benchmark examples The Smith and Jones arguments are **LUM.a$^{\mathbf{x}}$**-valid. Their premises are **SDL**-consistent and hence normal,

which means that (by Theorem 3.14), the adaptive logics are just as strong as **SDL** for these cases.[40] The Roberts and Thomas arguments are not valid in **LUM.a$^{\mathbf{r}}$** or **LUM.a$^{\mathbf{m}}$**. Here is a proof illustrating why the Roberts arguments are not valid in **LUM.a$^{\mathbf{x}}$**:

1	$\mathsf{O}(t \wedge r)$	Prem	$\emptyset$
2	$\mathsf{O}(\neg t \wedge v)$	Prem	$\emptyset$
3	$\mathsf{O}r$	1; RC	$\{\sharp(t \wedge r)\}\checkmark$
4	$\mathsf{O}v$	2; RC	$\{\sharp(\neg t \wedge v)\}\checkmark$
5	$\mathsf{O}(r \wedge v)$	3,4; RU	$\{\sharp(t \wedge r), \sharp(\neg t \wedge v)\}\checkmark$
6	$\sharp(t \wedge r) \vee \sharp(\neg t \wedge v)$	1,2; RU	$\emptyset$

In order to infer $\mathsf{O}r$ and $\mathsf{O}v$ via RC we need to rely on the falsity of $\sharp(t \wedge r)$ and $\sharp(\neg t \wedge v)$. However, further inspection of the premises teaches us that the disjunction of these abnormalities is **LUM.a**-derivable from the premises. To see why, note that this disjunction is **LUM.a**-equivalent to the following formula, which is a **LUM.a**-consequence of the premises:[41]

$$\begin{aligned}\big(\mathsf{O}(t \wedge r) \wedge \mathsf{O}\neg(t \wedge r)\big) \vee \big(\mathsf{O}(\neg t \wedge v) \wedge \mathsf{O}\neg(\neg t \wedge v)\big) \vee \\ (\mathsf{O}t \wedge \mathsf{O}\neg t) \vee (\mathsf{O}r \wedge \mathsf{O}\neg r) \vee (\mathsf{O}v \wedge \mathsf{O}\neg v)\end{aligned} \tag{19}$$

The minimal Dab-formula derived at line 6 blocks the derivation of the formulas derived at lines 3-5, causing the invalidity of the Roberts arguments. The same mechanism blocks the derivation of the conclusion of the Thomas argument.[42]

The Natascha argument, version 1, is **LUM.a**-valid (and hence **LUM.a$^{\mathbf{x}}$**-valid), but only because its premise set is **LUM.a**-trivial: from premises (i) and (ii) we can derive the negation of premise (iii) by (Agg). In contrast, the second version of the Natascha argument is **LUM.a**-satisfiable. Here is an **LUM.a$^{\mathbf{m}}$**-proof for this argument:

[40] Goble showed that the upper limit logic of **LUM.a$^{\mathbf{x}}$** is **SDL**, see [Goble, 2014, Observation 4.1].

[41] By **CL**, $\big(\mathsf{O}(t \wedge r) \wedge \mathsf{O}\neg(t \wedge r)\big) \vee \neg\big(\mathsf{O}(t \wedge r) \wedge \mathsf{O}\neg(t \wedge r)\big)$. Since $\mathsf{O}(t \wedge r)$, $\neg\big(\mathsf{O}(t \wedge r) \wedge \mathsf{O}\neg(t \wedge r)\big)$ entails $\mathsf{O}t$ by (RUM). Analogously, by **CL**, $\big(\mathsf{O}(\neg t \wedge v) \wedge \mathsf{O}\neg(\neg t \wedge v)\big) \vee \neg\big(\mathsf{O}(\neg t \wedge v) \wedge \mathsf{O}\neg(\neg t \wedge v)\big)$. Since $\mathsf{O}(\neg t \wedge v)$, $\neg\big(\mathsf{O}(\neg t \wedge v) \wedge \mathsf{O}\neg(\neg t \wedge v)\big)$ entails $\mathsf{O}\neg t$ by (RUM). Altogether, by **CL**, $\big(\mathsf{O}(t \wedge r) \wedge \mathsf{O}\neg(t \wedge r)\big) \vee \big(\mathsf{O}(\neg t \wedge v) \wedge \mathsf{O}\neg(\neg t \wedge v)\big) \vee (\mathsf{O}t \wedge \mathsf{O}\neg t)$. By **CL** again, (19) follows.

[42] In the Thomas case, the culpable Dab-formula is the disjunction $\sharp(t \wedge (f \vee s)) \vee \sharp(\neg t \wedge \neg f)$. We leave the verification to the interested reader.

1	$\mathsf{O}s$	Prem	$\emptyset$
2	$\mathsf{O}m$	Prem	$\emptyset$
3	$\mathsf{O}\neg(s \wedge m)$	Prem	$\emptyset$
4	$\mathsf{O}(s \supset t)$	Prem	$\emptyset$
5	$\mathsf{O}(m \supset t)$	Prem	$\emptyset$
6	$\mathsf{O}(s \wedge t)$	1, 4; RU	$\emptyset$
7	$\mathsf{O}(m \wedge t)$	2, 5; RU	$\emptyset$
8	$\mathsf{O}t$	6; RC	$\{\sharp(s \wedge t)\}$
9	$\mathsf{O}t$	7; RC	$\{\sharp(m \wedge t)\}$
10	$\sharp(s \wedge t) \vee \sharp(m \wedge t)$	1-3; RU	$\emptyset$

The formulas derived at lines 6 and 7 are **LUM.a**-derivable from the premises via applications of (Agg) and (RE). From each of these formulas we can derive $\mathsf{O}t$ via RC. Since we are working with the minimal abnormality strategy, lines 8 and 9 are unmarked at stage 10. If we were to use reliability, however, both lines would be marked. Indeed, the modified Natascha argument is valid for $\mathbf{LUM.a^m}$, while invalid for $\mathbf{LUM.a^r}$:

$$\mathsf{O}s, \mathsf{O}m, \mathsf{O}\neg(s \wedge m), \mathsf{O}(s \supset t), \mathsf{O}(m \supset t) \nvdash_{\mathbf{LUM.a^r}} \mathsf{O}t \tag{20}$$

$$\mathsf{O}s, \mathsf{O}m, \mathsf{O}\neg(s \wedge m), \mathsf{O}(s \supset t), \mathsf{O}(m \supset t) \vdash_{\mathbf{LUM.a^m}} \mathsf{O}t \tag{21}$$

The behavior of the ALs based on **LUM.b** and **LUM.c** is roughly analogous to the preceding case, with one exception. The premises in version 1 of the Natascha argument are inconsistent in **LUM.a** and **LUM.b**, but consistent in **LUM.c**. That is, we cannot aggregate premises (i) and (ii) of this argument, in the absence of the permission statements $\mathsf{P}s$ and $\mathsf{P}m$. Parallel to the situation for the modified Natascha argument in $\mathbf{LUM.a^x}$, we obtain the conclusion $\mathsf{O}t$ with $\mathbf{LUM.c^m}$ for the original Natascha argument, while we do not obtain it with $\mathbf{LUM.c^r}$. The following proof illustrates that $\mathsf{O}t$ is $\mathbf{LUM.c^m}$-derivable:[43]

[43]Lines 7 and 8 can be derived by means of (P-Agg). Note that this rule requires that the two formulas to be aggregated are themselves unconflicted. Hence we need RC to make these two derivations.

1	$\mathsf{O}s$	Prem	$\emptyset$
2	$\mathsf{O}m$	Prem	$\emptyset$
3	$\neg\mathsf{O}(s \wedge m)$	Prem	$\emptyset$
4	$\mathsf{O}(s \supset t)$	Prem	$\emptyset$
5	$\mathsf{O}(m \supset t)$	Prem	$\emptyset$
6	$\mathsf{O}(s \wedge m)$	1,2; RC	$\{\sharp s, \sharp m\}\checkmark^{12}$
7	$\mathsf{O}(s \wedge t)$	1, 4; RU	$\{\sharp s, \sharp(s \supset t)\}\checkmark^{12}$
8	$\mathsf{O}(m \wedge t)$	2, 5; RU	$\{\sharp m, \sharp(m \supset t)\}\checkmark^{12}$
9	$\mathsf{O}t$	7; RC	$\{\sharp s, \sharp(s \supset t), \sharp(s \wedge t)\}$
10	$\mathsf{O}t$	8; RC	$\{\sharp m, \sharp(m \supset t), \sharp(m \wedge t)\}$
11	$\sharp(s \wedge t) \vee \sharp(m \wedge t)$	1-3; RU	$\emptyset$
12	$\sharp s \vee \sharp m$	1-3; RU	$\emptyset$
13	$\sharp s \vee \sharp(m \wedge t)$	12; RU	$\emptyset$
14	$\sharp(s \wedge t) \vee \sharp m$	12; RU	$\emptyset$

The inferences at lines 13 and 14 hold in view of the **CL**-validity of $\sharp s \supset \sharp(s \wedge t)$ and $\sharp m \supset \sharp(m \wedge t)$ respectively. Where $\Gamma_n = \{\mathsf{O}s, \mathsf{O}m, \neg\mathsf{O}(s \wedge m), \mathsf{O}(s \supset t), \mathsf{O}(m \supset t)\}$:

$$\Phi_{14}(\Gamma_n) = \{\{\sharp(s \wedge t), \sharp s\}, \{\sharp(m \wedge t), \sharp m\}\} \tag{22}$$

It is easily verified that, in view of Definition 3.2, lines 9 and 10 are unmarked. If we were to use the reliability strategy instead, then by Definition 3.1 these lines would be marked in the proof above.

$$\Gamma_n \nvdash_{\mathbf{LUM.c^r}} \mathsf{O}t \tag{23}$$

$$\Gamma_n \vdash_{\mathbf{LUM.c^m}} \mathsf{O}t \tag{24}$$

The formula $\mathsf{O}t$ is a floating conclusion with respect to Γ_n. As pointed out in Section 4, it is a matter of debate whether or not floating conclusions are acceptable. We do not add anything to this debate here. It suffices for us to point out that each stance can be formally represented within the AL framework.

5.3 Further reading and open ends

The **LUM**-systems were introduced by Goble in [2004a; 2005; 2009], where they were called 'logics of permitted distribution' or **DPM**. They were called 'logics of unconflicted distribution' or **LUM** in [Goble, 2013; Goble, 2014]. Adaptive extensions of these systems were presented in [Straßer *et al.*, 2012; Straßer, 2014; Goble, 2014]. Moreover, in [Straßer, 2011], dyadic variants of the **LUM**-systems were also strengthened within the AL framework (see also Section 9.1 below).

There are many other types of deontic logics which invalidate (Inh). First, there is the general class of classical modal logics of which the logic **E** (cf. supra) is but one example. Second, Goble [1990a; 1990b] developed a very rich semantics for deontic logics, based on an idea from [Jackson, 1985]. On this semantics, $\mathsf{O}A$ is true iff the closest A-worlds are all better than the closest $\neg A$-worlds. Third and last, in more recent work, Cariani [2013] proposed yet another semantics for "ought" which invalidates (Inh) in a principled way – see also [Van De Putte, 2016; Van De Putte, 2019] for a formal investigation into this proposal. For each of these types of logics, one can ask whether it makes sense to strengthen them adaptively, and if so, which technical difficulties arise and what behavior the resulting logics will display. In particular, it would be interesting to learn whether some such variants perform better than the currently available logics, in dealing with the Roberts arguments and the Thomas argument.

6 Adaptive aggregation

A popular way to accommodate deontic conflicts in a formal system is by rejecting the aggregation principle (Agg), and with it the normality schema (K). In its simplest form, this proposal gives us the deontic logic **P**.[44] We will focus on two relatively basic ALs obtained from **P** in this section.

6.1 Adaptive aggregation: a basic example

Rejecting aggregation The language of **P** is the same as that of **SDL**, with P defined as the dual of O. As before, we will not consider nested occurrences of O. **P** is axiomatized by adding the axiom (D) to **CL** and closing the resulting set under modus ponens (MP), the necessitation rule (N), and the rule of inheritance (Inh). Each of the following are facts about the derivability relation of **P**:

[44] Again, we follow Goble's nomenclature. See the end of this section for pointers to the literature on this and related logics.

$$\vdash \mathsf{O}(p \vee \neg p) \tag{25}$$
$$\mathsf{O}p \vdash \mathsf{O}(p \vee q) \tag{26}$$
$$\mathsf{O}(p \wedge q) \vdash \mathsf{O}p, \mathsf{O}q \tag{27}$$
$$\mathsf{O}p, \mathsf{O}q \nvdash \mathsf{O}(p \wedge q) \tag{28}$$
$$\mathsf{O}(p \wedge (\neg p \vee q)) \vdash \mathsf{O}q \tag{29}$$
$$\mathsf{O}p, \mathsf{O}(\neg p \vee q) \nvdash \mathsf{O}q \tag{30}$$

In view of (Inh), Replacement of (Classical) Equivalents (RE) is valid in **P**. So in Chellas' terms, **P** is a non-normal but classical modal logic [Chellas, 1980].

One way to motivate and understand the rejection of (Agg) in **P** is in terms of multiple normative standards that ground our obligations, where $\mathsf{O}A$ is unspecific about the normative standard that grounds the obligation that A. Under such a reading, $\mathsf{O}A$ and $\mathsf{O}B$ may well be true even if there is no single standard that grounds the conjunction of both obligations, and hence $\mathsf{O}(A \wedge B)$ can still fail.[45] For instance, varying on our Smith example, one's duty to fight in the army might be based on the laws of one's country, whereas one's personal pacifist ethics grounds the claim that one ought not to fight in the army. Still, it does not follow that one ought to do the logically impossible, viz. to fight in the army and not fight in the army.

A semantics for **P** is obtained from the **SDL**-semantics (cf. Section 2) by generalizing the notion of an accessibility relation R. **P**-models are then of the type $\langle W, w_0, \mathcal{R}, V\rangle$, where W, w_0, and V are as before, but $\mathcal{R}$ is a non-empty *set* of serial accessibility relations, rather than a single such relation. The semantic clause for O then reads as follows:

(SC-O) $M, w \models \mathsf{O}A$ iff there is an $R \in \mathcal{R}$ such that $M, w' \models A$ for all w' such that Rww'

In other words, the single normative standard from **SDL** is replaced with a set of such standards, and we quantify (existentially) over such standards in order to determine the truth of $\mathsf{O}A$. It is well-known that **P** is sound and complete with respect to this semantics – see [Goble,

[45] The idea that one can relativize deontic logic to a given "moral code", and that what is obligatory under one such code may not be obligatory (or even forbidden) under another, is at least as old as Von Wright's *Deontic Logic* – see [von Wright, 1951, p. 15]. The difference here is that in **P**, the code that is at stake remains implicit, and $\mathsf{O}A$ only means that A is obligatory under at least *some* moral code.

2000, Theorem 1]. Other semantics can also be given for **P**. We refer the reader to [Goble, 2013, pp. 300-301] for an overview of these.

Going adaptive Even if aggregation is invalid on the reading of O just presented, in practice we do often aggregate our obligations. One simple way to argue for this is by referring to the benchmark examples from Section 4. It can easily be verified that neither the Smith argument nor the second variant of the Roberts argument is valid in **P**.

More generally, it is one thing to say that we take into account various normative standards and treat them as independent grounds or reasons when trying to determine what our obligations are. It is quite another thing to argue that none of these obligations can themselves be aggregated when doing so; this seems to go against much of our intuition.[46] For instance, when deciding how to get to my office in the morning, I may apply norms concerning the environment, norms uttered by my boss, and norms concerning my own safety and that of others. There seems to be no *prima facie* reason why we cannot integrate these various norms when settling for a single way to get to the office – e.g. I may conclude that I ought to bike to the office, since that way I will be in time for a meeting without causing air-pollution. The presence of deontic conflicts in itself seems insufficient to warrant a full rejection of aggregation, and, as we will show below, there is no logical reason for doing so either.

One needs to be careful here though. We cannot just add (Agg) to **P**, as this would give us again full **SDL** and hence deontic explosion in the face of deontic conflicts.[47] Moreover, as shown in [Goble, 2005, Sect. 2], there is no obvious conditional variant of (Agg) that can do a similar job, without in turn yielding some variant of deontic explosion.[48] So some obligations can be aggregated, but not all. As we will show in the remainder of this section, going adaptive allows us to steer a middle course between the weakness of **P** and deontic explosion.

The logics P$^{\mathbf{x}}$ The most straightforward way one might strengthen **P** adaptively, is by treating all formulas of the form $\mathsf{O}A \wedge \mathsf{O}B \wedge \neg\mathsf{O}(A \wedge B)$ as

[46] Compare [Goble, 2013, p. 253]: "Even if what one ought to do is often determined by different sources or authorities, insofar as propositions of what one ought to do serve as guides to action or as standards of evaluation of an agent's overall actions, there must be a common ought derived from those separate sources".

[47] We safely leave it to the reader to check that adding (Agg) to **P** yields full **SDL**.

[48] See also Section 5.2 of Goble's entry in the first volume of this handbook, [Goble, 2013]. In particular, Goble shows that adding the axiom (C-Agg) (cf. Section 5.1) to **P** will result in a variant of deontic explosion.

abnormalities. However, just as in the case of **ADPM1$^{\mathbf{r}}$**, this will give us a flip-flop. To see why, consider $\Gamma = \{\mathsf{O}p, \mathsf{O}\neg p, \mathsf{O}q, \mathsf{O}r\}$. Intuitively speaking, there is no problem with q and r in this example, and hence we expect $\mathsf{O}(q \wedge r)$ to be derivable. Such an inference can indeed be made within a proof of the adaptive logic thus defined. However, we can derive a disjunction of abnormalities (in that adaptive logic) from Γ which will block the derivation. This Dab-formula is a disjunction of the following three formulas:

$$\mathsf{O}q \wedge \mathsf{O}r \wedge \neg\mathsf{O}(q \wedge r)) \tag{31}$$

$$\mathsf{O}(p \vee \neg(q \wedge r)) \wedge \mathsf{O}\neg p \wedge \neg\mathsf{O}((p \vee \neg(q \wedge r)) \wedge \neg p) \tag{32}$$

$$\mathsf{O}(q \wedge r) \wedge \mathsf{O}\neg(q \wedge r) \wedge \neg\mathsf{O}((q \wedge r) \wedge \neg(q \wedge r)) \tag{33}$$

Suppose that (31) is false but the premises are true. Then $\mathsf{O}(q \wedge r)$ is the case. Likewise, since $\mathsf{O}(p \vee \neg(q \wedge r))$ follows by (Inh) from $\mathsf{O}p$, (32) can only be false (in view of the premises) if its last conjunct is false, and hence $\mathsf{O}\neg(q \wedge r)$ is true. But then the third abnormality, (33) must be true.

It is not hard to see where the problem could be in cases like this. That is, since $\mathsf{O}p, \mathsf{O}\neg p \in \Gamma$, we should not use these obligations – nor weakenings of them – in order to apply aggregation. In other words, obligations that are themselves conflicted, or subformulas of which are conflicted, should be treated as abnormal.

This brings us to a slightly more complicated set of abnormalities, which is due to Goble [2014]. As before, let $\sharp(A)$ denote the disjunction of all formulas $\mathsf{O}B \wedge \mathsf{O}\neg B$, where $B \in S(A)$ (B is a subformula of A). Let $\natural(A, B) = (\mathsf{O}A \wedge \mathsf{O}B \wedge \neg\mathsf{O}(A \wedge B)) \vee \sharp(A \wedge B)$. We now define

$$\Omega_{\mathbf{P}} = \{(\natural(A, B) \mid A, B \in \mathcal{W}\}$$

In other words, we have an abnormality with respect to A and B iff they are both obligatory and their conjunction is not obligatory, *or* a proper subformula of them is conflicted. This means that as soon as e.g. $\mathsf{O}p, \mathsf{O}\neg p$ holds, all abnormalities $\natural(A, B)$ with $p \in S(A)$ are true. Under this definition, none of the formulas (31)-(33) are abnormalities. The corresponding disjunction of $\Omega_{\mathbf{P}}$-abnormalities

$$\natural(q, r) \vee \natural(p \vee \neg(q \wedge r), \neg p) \vee \natural(q \wedge r, \neg(q \wedge r)) \tag{34}$$

is not a minimal Dab-consequence of Γ, since $\natural(p \vee \neg(q \wedge r), \neg p)$ alone follows from Γ.

Let the logics $\mathbf{P^r}$ and $\mathbf{P^m}$ be the adaptive logics defined by the triple $\langle \mathbf{P}, \Omega_{\mathbf{P}}, \mathsf{x} \rangle$, where $\mathsf{x} \in \{\mathsf{r}, \mathsf{m}\}$.[49] It can easily be checked that the upper limit logic of $\mathbf{P^r}$ and $\mathbf{P^m}$ is just **SDL**: adding the negation of all members of $\Omega_{\mathbf{P}}$ as axioms to $\mathbf{P}$, is equivalent to adding (Agg) to $\mathbf{P}$.[50] This means that normal premise sets in the logics $\mathbf{P^x}$ are just **SDL**-consistent premise sets (where 'normal' is understood in the technical sense specified on page 392). Hence by Theorem 3.14, whenever a premise set is **SDL**-consistent, its $\mathbf{P^x}$-consequence set will be identical to its **SDL**-consequence set:

Theorem 6.1. *If* Γ *is* **SDL***-consistent, then* $\mathsf{Cn}_{\mathbf{P^r}}(\Gamma) = \mathsf{Cn}_{\mathbf{P^m}}(\Gamma) = \mathsf{Cn}_{\mathbf{SDL}}(\Gamma)$.

6.2 Evaluating the logics

Explosion principles The logic $\mathbf{P}$, and with it $\mathbf{P^r}$ and $\mathbf{P^m}$, clearly accommodates conflicts of the basic type $\mathsf{O}A, \mathsf{O}\neg A$. By (RE), ($\mathsf{Def_P}$) and **CL**-properties, also conflicts of the type $\mathsf{O}A, \neg\mathsf{P}A$ are consistent in $\mathbf{P}$ and its adaptive extensions.

All other types of deontic conflicts listed in Section 4.2 will be trivialized within these logics. The reasons are similar to those for **LUM.b** and **LUM.c**: $\mathsf{O}(A \wedge \neg A)$ is contradictory in view of (P), $\mathsf{O}A \wedge \mathsf{P}\neg A$ is contradictory in view of ($\mathsf{Def_P}$) and (RE), and $\mathsf{P}(A \wedge \neg A)$ is false in view of (N) and ($\mathsf{Def_P}$). So the simplicity of $\mathbf{P}$ comes at an important price, viz. that it can only handle conflicting obligations and does not allow us to reason about conflicting information concerning (obligations and) permissions.[51]

Benchmark examples The arguments for Jones and Roberts 1 are valid in both $\mathbf{P^r}$ and $\mathbf{P^m}$. This is easy to verify since the arguments are already valid in $\mathbf{P}$ in view of its validating (Inh), and since both $\mathbf{P^r}$ and $\mathbf{P^m}$ are extensions of $\mathbf{P}$. The premises of the Smith argument are

[49] In Goble's work, the first of these two logics is known as $\mathbf{AP^r}$. As before, we skip the initial "A" since the superscript suffices to mark the difference with the monotonic logic $\mathbf{P}$.

[50] To see why this is so, note first that if we negate all formulas of the form $\natural(A, B)$, then *a fortiori* we negate all formulas of the form $\mathsf{O}A \wedge \mathsf{O}B \wedge \neg\mathsf{O}(A \wedge B)$, and hence we affirm all instances of (Agg). In addition, we also negate all formulas of the form $\mathsf{O}A \wedge \mathsf{O}\neg A$, but these are anyway **SDL**-valid.

[51] Note also that simply rejecting (P) will not allow us to have a satisfactory account of conflicts of the type $\mathsf{O}(A \wedge \neg A)$: due to (Inh) these conflicts will still lead to deontic explosion.

normal: no Dab-formula can be derived from them. As a result, we can aggregate the obligations in the argument and derive $\mathsf{O}s$.

The second Roberts argument is also valid in $\mathbf{P^x}$, but here the reasoning is slightly more intricate. First, applying (Inh), we can derive $\mathsf{O}r$ and $\mathsf{O}v$ from the premises. To apply aggregation to these two formulas, we need to assume that neither r nor v are conflicted, given the premise set. This is clearly the case: the only conflict that follows from the premises, is $\mathsf{O}t, \mathsf{O}\neg t$. The following $\mathbf{P^r}$-proof illustrates how we can obtain the desired conclusion for Roberts 2, while avoiding the aggregation of conflicted obligations:

1	$\mathsf{O}(t \wedge r)$	Prem	$\emptyset$
2	$\mathsf{O}(\neg t \wedge v)$	Prem	$\emptyset$
3	$\mathsf{O}r$	1; RU	$\emptyset$
4	$\mathsf{O}v$	2; RU	$\emptyset$
5	$\mathsf{O}(r \wedge v)$	4,5; RC	$\{\natural(r, v)\}$
6	$\mathsf{O}t$	1; RU	$\emptyset$
7	$\mathsf{O}\neg t$	2; RU	$\emptyset$
8	$\mathsf{O}(t \wedge v)$	6,4; RU	$\{\natural(t, v)\}\checkmark^{11}$
9	$\mathsf{O}(t \wedge \neg t)$	6,7; RC	$\{\natural(t, \neg t)\}\checkmark^{10}$
10	$\natural(t, \neg t)$	6,7; RU	$\emptyset$
11	$\natural(t, v)$	6,7; RU	$\emptyset$

Since $\natural(t, v)$ follows from the premises, we cannot finally derive $\mathsf{O}(t \wedge v)$ from them. So even if there is no direct conflict between t and v, the fact that t is itself conflicted is sufficient to block its aggregation with other (unproblematic) obligations.[52]

The reasoning for the Thomas argument is wholly analogous to the second Roberts case, with the difference that we apply (Inh) once more after aggregating $\mathsf{O}(f \vee s)$ and $\mathsf{O}\neg f$ to $\mathsf{O}((f \vee s) \wedge \neg s)$. This gives us the desired conclusion $\mathsf{O}s$.

For the Natascha arguments, it turns out that with the $\mathbf{P}$-based adaptive logics the strategies make no difference. The point is that, although we can obviously not apply aggregation to $\mathsf{O}s$ and $\mathsf{O}m$, we can still aggregate $\mathsf{O}s$ and $\mathsf{O}(s \supset t)$ (and likewise, $\mathsf{O}m$ and $\mathsf{O}(m \supset t)$). The fact that the pair (m, s) behaves abnormally ($\natural(m, s)$ follows from the premises of the argument) does not imply that either of $(s, s \supset t)$ or $(m, m \supset t)$ behave abnormally. Hence we can finally derive $\mathsf{O}t$ on two different conditions in both $\mathbf{P^r}$ and $\mathbf{P^m}$. We illustrate this for the first

[52] As pointed out by Goble, allowing aggregation for all A, B such that $A \wedge B$ is consistent is simply a no-go in the context of $\mathbf{P}$, since it will lead to another form of deontic explosion. See [Goble, 2005, Sect. 2.4.1].

variant of the Natascha argument:

1	$\mathsf{O}s$	Prem	$\emptyset$
2	$\mathsf{O}m$	Prem	$\emptyset$
3	$\neg\mathsf{O}(s \wedge m)$	Prem	$\emptyset$
4	$\mathsf{O}(s \supset t)$	Prem	$\emptyset$
5	$\mathsf{O}(m \supset t)$	Prem	$\emptyset$
6	$\mathsf{O}(s \wedge (s \supset t))$	1,4; RC	$\{\natural(s, s \supset t)\}$
7	$\mathsf{O}(m \wedge (m \supset t))$	2, 5; RC	$\{\natural(m, m \supset t)\}$
8	$\mathsf{O}t$	6; RU	$\{\natural(s, s \supset t)\}$
9	$\mathsf{O}t$	7; RU	$\{\natural(m, m \supset t)\}$

For none of the variants of the Natascha arguments, the disjunction of abnormalities $(\mathsf{O}s \wedge \mathsf{O}\neg s) \vee (\mathsf{O}m \wedge \mathsf{O}\neg m)$ is **P**-derivable from the premises. Nor is there another Dab-formula which prevents lines 8 and 9 from being finally derivable. So, to sum up, all inferences from our benchmark examples are valid in the logics $\mathbf{P^r}$ and $\mathbf{P^m}$.

This is not to say that there is no difference between $\mathbf{P^r}$ and $\mathbf{P^m}$. Consider e.g. $\Gamma = \{\mathsf{O}p, \mathsf{O}q, \mathsf{O}r, \neg\mathsf{O}(p \wedge r) \vee \neg\mathsf{O}(q \wedge r)\}$. From this premise set, $\mathbf{P^r}$ will not allow us to finally derive $\mathsf{O}(p \wedge r) \vee \mathsf{O}(q \wedge r)$, whereas $\mathbf{P^m}$ will. To understand this, note that $\natural(p, r) \vee \natural(q, r)$ is a minimal Dab-consequence of Γ, whence both abnormalities are unreliable in view of Γ. However, since nothing prevents us to assume that either the first or the second abnormality is *false*, using *minimal abnormality* we can derive $\mathsf{O}(p \wedge r) \vee \mathsf{O}(q \wedge r)$.

6.3 Further reading and open ends

Bernard Williams [Williams and Atkinson, 1965] was the first to advocate a rejection of (Agg) on philosophical grounds; Marcus [1980] is another important proponent of such a rejection. More formally worked out proposals can be found in [van Fraassen, 1973; Chellas, 1980; Schotch and Jennings, 1981]. Later, Goble developed the semantics and metatheory of **P** and variants of it in detail – see in particular [Goble, 2000; Goble, 2004b; Goble, 2003]. For a more complete overview of the literature on **P** and close (monotonic) relatives, we refer to Section 5.2 of Goble's entry [Goble, 2013] in the first volume of this handbook.

The first adaptive logic that applies the idea of "adaptive aggregation" was published in [Meheus *et al.*, 2010], and later reworked in [Meheus *et al.*, 2012]. These logics are however based on a richer lower limit logic, viz. the logic $\mathbf{SDL}_a\mathbf{P}_e$ from [Goble, 2000]. In this system, one can express both an "existential" notion of obligation O_e (whose logic is

P) and a "universal" notion of obligation O_a, whose logic is **SDL**. The two modalities are connected by the following bridging principle:

(B) $\mathsf{O}_a(A \supset B) \wedge \mathsf{O}_e A \vdash \mathsf{O}_e B$

which entails i.a. that every universal obligation is also an existential obligation, $\mathsf{O}_a A \supset \mathsf{O}_e A$. Alternatively, one can interpret the logics in terms of our distinction between *prima facie* obligations and actual obligations (cf. Section 3.1).

Adaptive logics that are based on **P** itself are discussed in [Goble, 2014]; here we only discussed the second of the two. The other AL discussed by Goble appears to be slightly weaker. For instance, in this logic, the Natascha argument is only valid if we use *minimal abnormality*. More generally, in this logic any conflict of the type $\mathsf{O}A \wedge \mathsf{O}B \wedge \neg\mathsf{O}(A \wedge B)$ "infects" all the subformulas of A and B. We leave the full inspection and proof of this claim for another occasion.

An interesting issue concerns the enrichment of the aforementioned ALs with operators that allow one to express (technical, physical, practical) impossibility at the object level. Indeed, in Williams' famous essay, he argues that purely logical conflicts between oughts are only a special case of a much more common type of conflicts, viz. conflicts between two obligations whose joint fulfillment is impossible for *contingent* reasons – e.g. because of the particular physical situation we find ourselves in [Williams and Atkinson, 1965]. This raises a number of questions concerning the interplay between alethic and deontic modalities, which would take us well beyond the scope of the present chapter – see however [Beirlaen, 2012, Chapter 4] for a first attempt to combine alethic and deontic modalities.

7 Inconsistency-adaptive deontic logics

As noted in Section 3.5, the first adaptive logics were *inconsistency-adaptive*. Inconsistency-adaptive logics are members of the larger family of *paraconsistent logics*, i.e. logics which invalidate (ECQ).

Note that (ECQ) bears close affinity to (DEX). To obtain the latter from the former we only need to prefix the formulas involved with an O-operator. Besides the approaches we saw in Sections 5 and 6, a third natural way to invalidate (DEX) is by invalidating (ECQ).

Going paraconsistent has a couple of additional benefits in the context of deontic logic. A first is that it allows us to preserve the interdefinability of O and P, while invalidating (DEX-$\mathsf{OP}\neg$). Assuming the

interdefinability of O and P, the formula $\mathsf{O}A \wedge \mathsf{P}\neg A$ is equivalent to the contradictions $\mathsf{O}A \wedge \neg\mathsf{O}A$ and $\neg\mathsf{P}\neg A \wedge \mathsf{P}\neg A$. By (ECQ), these contradictions entail everything. To prevent such explosive behavior, it suffices to invalidate (ECQ).

A second advantage is that only a paraconsistent deontic logic can invalidate the explosion principles (DEX-$\mathsf{O}\neg\mathsf{O}$) and (DEX-$\mathsf{P}\neg\mathsf{P}$), for the obvious reason that these principles are instances of (ECQ):

$$\mathsf{O}A, \neg\mathsf{O}A \vdash \mathsf{O}B \qquad \text{(DEX-O}\neg\text{O)}$$
$$\mathsf{P}A, \neg\mathsf{P}A \vdash \mathsf{O}B \qquad \text{(DEX-P}\neg\text{P)}$$

There are independent reasons as to why, in some contexts, we may want to tolerate *contradictory norms*, i.e. formulas of the form $\mathsf{O}A \wedge \neg\mathsf{O}A$ or $\mathsf{P}A \wedge \neg\mathsf{P}A$. Priest, for instance, gives the following example. Suppose that, in some country, women are not permitted to vote, while property holders are permitted to vote. Suppose further that, perhaps due to a recent revision of the property law, women are permitted to hold property. Then female property holders are both permitted and not permitted to vote ($\mathsf{P}v \wedge \neg\mathsf{P}v$) [Priest, 1987, pp. 184–185].

In this section, we present inconsistency-adaptive deontic logics. We will work stepwise, starting with the paraconsistent logic **CLuN**, its deontic extension **DCLuN**, and adaptive strengthenings **DCLuN$^{\mathbf{x}}$** (Section 7.1). After that, we will consider several variants of **DCLuN** and their associated adaptive logics (sections 7.2 and 7.3).

7.1 Paraconsistent adaptive deontic logic

A paraconsistent core logic We use the paraconsistent logic **CLuN** as our starting point. **CLuN** is an acronym for 'Classical Logic with gluts for Negation'. A truth-value *glut* for negation relative to a formula A occurs when both A and its negation are true; **CLuN** allows such gluts whereas **CL** disallows them. The deontic logics to be presented in this section are extensions of **CLuN**, but they are defined so that plenty of other paraconsistent logics may replace **CLuN** as their core logic. In Sections 7.2 and 7.3 we will mention some alternatives.

The set $\mathcal{W}^\sim$ of well-formed **CLuN**-formulas is the following:

$$\mathcal{W}^\sim := \mathcal{S} \mid \sim\langle\mathcal{W}^\sim\rangle \mid \neg\langle\mathcal{W}^\sim\rangle \mid \langle\mathcal{W}^\sim\rangle \vee \langle\mathcal{W}^\sim\rangle \mid \langle\mathcal{W}^\sim\rangle \wedge \langle\mathcal{W}^\sim\rangle \mid \langle\mathcal{W}^\sim\rangle \supset \langle\mathcal{W}^\sim\rangle \mid \langle\mathcal{W}^\sim\rangle \equiv \langle\mathcal{W}^\sim\rangle$$

In the remainder, we will stick to $\neg$ as the connective denoting classical negation. Beside $\neg$, $\mathcal{W}^\sim$ contains the connective $\sim$ which we will use as our paraconsistent negation sign. In fact, $\sim$ is the only **CLuN**-connective which behaves differently from the classical connectives. We

obtain **CLuN** by adding the following axiom schema to **CL**:

$$A \vee {\sim}A \tag{EM$\sim$}$$

We write $\Gamma \vdash_{\mathbf{CLuN}} A$ to denote that A is **CLuN**-derivable from Γ.

The **CLuN**-semantics is defined as follows. To obtain a **CLuN**-model M, we extend the assignment function v_a of **CL** so that it assigns truth values not only to schematic letters, but also to formulas of the form ${\sim}A$, i.e. $v_a : \mathcal{S} \cup \{{\sim}A \mid A \in \mathcal{W}^{\sim}\} \rightarrow \{0, 1\}$. Next, we extend v_a to a valuation function v as follows:

(SC1) For formulas $A \in \mathcal{S} \cup \{{\sim}A \mid A \in \mathcal{W}^{\sim}\} : M \models A$ iff $v_a(A) = 1$.
(SC2) For $\neg, \vee, \wedge, \supset, \equiv$, the semantic clauses for **CLuN** are those of **CL**.

Finally, in order to validate the axiom (EM$\sim$), we require that all **CLuN**-models satisfy the following condition: for all $A \in \mathcal{W}^{\sim}$, $M \models A$ or $M \models {\sim}A$. A semantic consequence relation for **CLuN** is defined as follows: $\Gamma \Vdash_{\mathbf{CLuN}} A$ iff for all **CLuN**-models M: if $M \models B$ for all $B \in \Gamma$, then $M \models A$.

Before we move on to deontic extensions of **CLuN**, we point out a number of relevant properties of this logic for ease of reference:

(i) **CLuN** is paraconsistent, but not *paracomplete*: while (ECQ) is **CLuN**-invalid for $\sim$, the excluded middle principle (EM$\sim$) is **CLuN**-valid.

(ii) In contrast to well-known paraconsistent logics such as Priest's **LP**, **CLuN** validates modus ponens:

$$A, A \supset B \vdash B \tag{MP}$$

Note that $A \supset B$ and ${\sim}A \vee B$ are not **CLuN**-equivalent: if $v(A) = v({\sim}A) = v({\sim}B) = 1$ and $v(B) = 0$, then $v(A \supset B) = 0$ while $v({\sim}A \vee B) = 1$.

(iii) De Morgan's laws and the double negation laws are invalid for $\sim$ in **CLuN**. This means that complex contradictions are not reducible to contradictions between elementary letters:

$$(p \wedge q) \wedge {\sim}(p \wedge q) \nvdash (p \wedge {\sim}p) \vee (q \wedge {\sim}q) \tag{35}$$
$$(p \vee q) \wedge {\sim}(p \vee q) \nvdash (p \wedge {\sim}p) \vee (q \wedge {\sim}q) \tag{36}$$
$$(p \supset q) \wedge {\sim}(p \supset q) \nvdash (p \wedge {\sim}p) \vee (q \wedge {\sim}q) \tag{37}$$
$${\sim}{\sim}(p \wedge {\sim}p) \nvdash p \wedge {\sim}p \tag{38}$$

(iv) Contraposition, modus tollens, and disjunctive syllogism are invalid for $\sim$ in **CLuN**:

$$A \supset B \nvdash {\sim}B \supset {\sim}A \tag{39}$$
$$A \supset B, {\sim}B \nvdash {\sim}A \tag{40}$$
$$A \vee B, {\sim}A \nvdash B \tag{41}$$

A paraconsistent deontic logic A technically straightforward way to construct a deontic logic on the basis of **CLuN** is the following. First, we extend the language $\mathcal{W}^{\sim}$ with the deontic operator O, preventing nested occurrences of the deontic operator:

$$\mathcal{W}^{\sim}_{\mathsf{O}} := \quad \mathcal{W}^{\sim} \mid \mathsf{O}\langle\mathcal{W}^{\sim}\rangle \mid {\sim}\langle\mathcal{W}^{\sim}_{\mathsf{O}}\rangle \mid \neg\langle\mathcal{W}^{\sim}_{\mathsf{O}}\rangle \mid \langle\mathcal{W}^{\sim}_{\mathsf{O}}\rangle \vee \langle\mathcal{W}^{\sim}_{\mathsf{O}}\rangle \mid \langle\mathcal{W}^{\sim}_{\mathsf{O}}\rangle \wedge \langle\mathcal{W}^{\sim}_{\mathsf{O}}\rangle \mid \langle\mathcal{W}^{\sim}_{\mathsf{O}}\rangle \supset \langle\mathcal{W}^{\sim}_{\mathsf{O}}\rangle \mid \langle\mathcal{W}^{\sim}_{\mathsf{O}}\rangle \equiv \langle\mathcal{W}^{\sim}_{\mathsf{O}}\rangle$$

The logic **DCLuN** is axiomatized by adding to **CLuN** the axioms (K), (D), and closing the resulting set under (N) and (MP). Note that for (D) we need the original version (cf. page 376), hence with classical negations ($\neg$) only.

The semantics for **DCLuN** looks as follows. A model is a quadruple $M = \langle W, w_0, R, v\rangle$ where W is a non-empty set, $w_0 \in W$, $R \subseteq W \times W$ is a serial accessibility relation, and $v : \mathcal{W}^{\sim}_{\mathsf{O}} \times W \to \{1, 0\}$ is a valuation function. As with **CLuN**, we first assign truth values to both schematic letters and formulas of the form ${\sim}A$: $v_a : \mathcal{S} \cup \{{\sim}A \mid A \in \mathcal{W}^{\sim}_{\mathsf{O}}\} \times W \to \{0, 1\}$. v_a is extended to v as follows:

(SC1') For formulas $A \in \mathcal{S} \cup \{{\sim}A \mid A \in \mathcal{W}^{\sim}_{\mathsf{O}}\}$: $M, w \models A$ iff $v_a(A, w) = 1$.

(SC2') For $\mathsf{O}, \neg, \vee, \wedge, \supset, \equiv$, the semantic clauses for **DCLuN** are exactly those of **SDL** (cf. Section 2).

A model M is a **DCLuN**-*model* iff it satisfies the following condition on v:

$$\text{for all } w \in W, \text{ for all } A : v(A, w) = 1 \text{ or } v({\sim}A, w) = 1 \tag{C_u}$$

$\Gamma \Vdash_{\mathbf{DCLuN}} A$ iff for all **DCLuN**-models M: if $M, w_0 \models B$ for all $B \in \Gamma$, then $M, w_0 \models A$.

The proof of soundness for this logic is a matter of routine. For completeness, we can use the well-known technique of canonical models (see e.g. [Blackburn *et al.*, 2001, Chapter 4]), adjusted to the setting with an actual world. Fix a maximal, $\neg$-consistent set $\Gamma \subseteq \mathcal{W}^{\sim}_{\mathsf{O}}$. We build the canonical model $M^c_\Gamma = \langle W^c, \Gamma, R^c, V^c\rangle$ for this set as follows:

(i) W^c is the set of all maximal consistent and **DCLuN**-closed sets Δ,

(ii) $R^c = \{(\Delta, \Delta') \mid \{A \mid \mathsf{O}A \in \Delta\} \subseteq \Delta'\}$,

(iii) for all $A \in \mathcal{S} \cup \{\sim A \mid A \in \mathcal{W}^{\sim}_{\mathsf{O}}\}$, for all $\Delta \in W^c$: $v_a(A, \Delta) = 1$ iff $A \in \Delta$.

To show that M^c_Γ is a **DCLuN**-model, we need to rely on excluded middle for $\sim$ and the maximality of each $\Delta \in W^c$. For seriality, we rely on the (D)-axiom in the usual way. The proof of the truth lemma proceeds by a standard induction. So we can derive that all the members of Γ are satisfied at Γ in M^c_Γ.

Note that, since **CLuN** is a conservative extension of **CL**, **DCLuN** is also a conservative extension of **SDL**. However, if we consider the $\neg$-free fragment of **DCLuN**, and treat $\sim$ as the "proper" negation, then **DCLuN** is a proper fragment of **SDL**. When applying the logic **DCLuN** to concrete examples, we will use $\sim$ to translate negations in natural language. Given this convention, the logic **DCLuN** is strongly conflict-tolerant.

$$\mathsf{O}A \wedge \mathsf{O}{\sim}A \nvdash_{\mathbf{DCLuN}} \mathsf{O}B$$
$$\mathsf{O}A \wedge {\sim}\mathsf{O}A \nvdash_{\mathbf{DCLuN}} \mathsf{O}B$$

In **DCLuN** we can define permission in various ways relative to our negation operators:

$$\mathsf{P}^{\neg}_{\neg}A =_{\mathsf{df}} \neg\mathsf{O}\neg A$$
$$\mathsf{P}^{\neg}_{\sim}A =_{\mathsf{df}} \neg\mathsf{O}{\sim}A$$
$$\mathsf{P}^{\sim}_{\neg}A =_{\mathsf{df}} {\sim}\mathsf{O}\neg A$$
$$\mathsf{P}^{\sim}_{\sim}A =_{\mathsf{df}} {\sim}\mathsf{O}{\sim}A$$

All of these permission operators tolerate conflicts between an obligation and a permission, as well as contradictory norms. Where $\dagger, \ddagger \in \{\sim, \neg\}$:

$$\mathsf{O}A \wedge \mathsf{P}^{\ddagger}_{\dagger}{\sim}A \nvdash_{\mathbf{DCLuN}} \mathsf{O}B \tag{42}$$
$$\mathsf{O}{\sim}A \wedge \mathsf{P}^{\ddagger}_{\dagger}A \nvdash_{\mathbf{DCLuN}} \mathsf{O}B \tag{43}$$
$$\mathsf{P}^{\ddagger}_{\dagger}A \wedge {\sim}\mathsf{P}^{\ddagger}_{\dagger}A \nvdash_{\mathbf{DCLuN}} \mathsf{O}B \tag{44}$$

In sum, **DCLuN** is very conflict-tolerant, especially compared to the logics discussed in previous sections. However, it is also rather weak. To be sure, the Jones argument, the Roberts arguments, and the (original

and modified) Natascha argument are valid in **DCLuN** due to the validity of (Inh) and (Agg). Unfortunately, the Smith argument and the Thomas argument are not **DCLuN**-valid. More generally, all instances of the following inference schemas fail in **DCLuN**:

$$\mathsf{O}(A \supset B) \nvdash_{\mathbf{DCLuN}} \mathsf{O}(\sim B \supset \sim A) \tag{45}$$

$$\mathsf{O}(A \supset B), \mathsf{O}{\sim}B \nvdash_{\mathbf{DCLuN}} \mathsf{O}{\sim}A \tag{46}$$

$$\mathsf{O}(A \vee B), \mathsf{O}{\sim}A \nvdash_{\mathbf{DCLuN}} \mathsf{O}B \tag{47}$$

The invalidity of (45)-(47) mirrors the invalidity of their non-deontic counterparts (39)-(41) in **CLuN**. So the main advantage of **DCLuN** goes hand in hand with its inability to validate seemingly intuitive inferences. This drawback is overcome by strengthening this system within the adaptive logics framework.

Going adaptive We strengthen **DCLuN** to the adaptive logic **DCLuN$^{\mathsf{x}}$**, which is defined by the triple $\langle \mathbf{DCLuN}, \Omega^{\sim}, \mathsf{x} \rangle$, where

$$\Omega^{\sim} = \{A \wedge \sim A \mid A \in \mathcal{W}^{\sim}_{\mathsf{O}}\} \cup \{\mathsf{P}^{\neg}_{\neg}(A \wedge \sim A) \mid A \in \mathcal{W}^{\sim}\}$$

$\Omega^{\sim}$ contains not only plain contradictions, but also formulas that express that in some deontically accessible world, a given contradiction is true. This allows us at once to validate the Smith argument and the Thomas argument. Here is a **DCLuN$^{\mathsf{x}}$**-proof illustrating the validity of the Thomas argument:

1	$\mathsf{O}(t \wedge (f \vee s))$	Prem	$\emptyset$
2	$\mathsf{O}(\sim t \wedge \sim f)$	Prem	$\emptyset$
3	$\mathsf{O}(f \vee s)$	1; RU	$\emptyset$
4	$\mathsf{O}{\sim}f$	2; RU	$\emptyset$
5	$\mathsf{O}s$	3,4; RC	$\{\mathsf{P}^{\neg}_{\neg}(f \wedge \sim f)\}$

The inference made at line 5 holds in view of the **DCLuN**-valid inference

$$\mathsf{O}(f \vee s), \mathsf{O}{\sim}f \vdash \mathsf{O}s \vee \mathsf{P}^{\neg}_{\neg}(f \wedge \sim f) \tag{48}$$

Suppose that $\mathsf{O}(f \vee s)$ and $\mathsf{O}{\sim}f$. By (Agg), $\mathsf{O}((f \vee s) \wedge \sim f)$. By normal modal logic properties, we can infer $\mathsf{O}s \vee \neg\mathsf{O}\neg(f \wedge \sim f)$ so that we can derive $\mathsf{O}s$ on the condition $\mathsf{P}^{\neg}_{\neg}(f \wedge \sim f)$.

Equations (49)-(54) illustrate that the **DCLuN**-invalid inferences (39)-(41) and (45)-(47) hold conditionally in **DCLuN$^{\mathsf{x}}$**. The conditions

on which these inferences can be made in a **DCLuN$^{\mathbf{x}}$**-proof are indicated between square brackets.

$$p \supset q \vdash_{\mathbf{DCLuN^x}} {\sim}q \supset {\sim}p \qquad [q \wedge {\sim}q] \qquad (49)$$

$$p \supset q, {\sim}q \vdash_{\mathbf{DCLuN^x}} {\sim}p \qquad [q \wedge {\sim}q] \qquad (50)$$

$$p \vee q, {\sim}p \vdash_{\mathbf{DCLuN^x}} q \qquad [p \wedge {\sim}p] \qquad (51)$$

$$\mathsf{O}(p \supset q) \vdash_{\mathbf{DCLuN^x}} \mathsf{O}({\sim}q \supset {\sim}p) \qquad [\mathsf{P}^{\neg}_{\neg}(q \wedge {\sim}q)] \qquad (52)$$

$$\mathsf{O}(p \supset q), \mathsf{O}{\sim}q \vdash_{\mathbf{DCLuN^x}} \mathsf{O}{\sim}p \qquad [\mathsf{P}^{\neg}_{\neg}(q \wedge {\sim}q)] \qquad (53)$$

$$\mathsf{O}(p \vee q), \mathsf{O}{\sim}p \vdash_{\mathbf{DCLuN^x}} \mathsf{O}q \qquad [\mathsf{P}^{\neg}_{\neg}(p \wedge {\sim}p)] \qquad (54)$$

More generally, relative to premise sets from which no abnormalities are **DCLuN**-derivable $\sim$ is as strong as $\neg$ in **DCLuN$^{\mathbf{x}}$**. That is, where $A \in \mathcal{W}^{\sim}_{\mathsf{O}}$, let $\pi(A)$ be the result of replacing every occurrence of $\sim$ in A with $\neg$. We lift this translation to sets of formulas in the usual way. We can now prove the following:

Theorem 7.1. *If* Γ *is normal, then* $\Gamma \vdash_{\mathbf{DCLuN^x}} A$ *iff* $\pi(\Gamma) \vdash_{\mathbf{SDL}} \pi(A)$.

Proof. The upper limit logic of **DCLuN$^{\mathbf{x}}$** is obtained by adding to **DCLuN** all formulas $\neg A$ for which $A \in \Omega^{\sim}$. Call this logic **UDCLuN**. By Theorem 3.14: If Γ is normal, then $\Gamma \vdash_{\mathbf{DCLuN^x}} A$ iff $\Gamma \vdash_{\mathbf{UDCLuN}} A$. We show that $\Gamma \vdash_{\mathbf{UDCLuN}} A$ iff $\pi(\Gamma) \vdash_{\mathbf{SDL}} \pi(A)$.
($\Rightarrow$) It is easily checked that, under the transformation given, all **CLuN**-valid inferences are **CL**-valid; (K), (D), and (N) are **SDL**-valid; and all elements of $\pi(\{\neg A \mid A \in \Omega^{\sim}\})$ are **SDL**-valid.
($\Leftarrow$) Given the fact that **UDCLuN**, like **DCLuN**, extends **SDL**, it suffices to show that $\sim$ is as strong as $\neg$ in **UDCLuN**:

$$\vdash_{\mathbf{UDCLuN}} {\sim}A \supset \neg A \qquad (55)$$

$$\vdash_{\mathbf{UDCLuN}} \mathsf{O}{\sim}A \supset \mathsf{O}\neg A \qquad (56)$$

Ad. (55) Suppose $\sim A$. Then $\neg A \vee (A \wedge {\sim}A)$ since $\vdash_{\mathbf{CLuN}} {\sim}A \supset (\neg A \vee (A \wedge {\sim}A))$. We also know that $\vdash_{\mathbf{UDCLuN}} \neg(A \wedge {\sim}A)$, so by **CL**-properties we obtain $\neg A$.
*Ad. (56)*By (N), $\vdash_{\mathbf{UDCLuN}} \mathsf{O}({\sim}A \supset (\neg A \vee (A \wedge {\sim}A)))$. Suppose $\mathsf{O}{\sim}A$. By (K) and (MP), $\mathsf{O}(\neg A \vee (A \wedge {\sim}A))$. By **SDL**-properties, $\mathsf{O}\neg A \vee \mathsf{P}^{\neg}_{\neg}(A \wedge {\sim}A)$. But then $\mathsf{O}\neg A$ follows in view of $\vdash_{\mathbf{UDCLuN}} \neg\mathsf{P}^{\neg}_{\neg}(A \wedge {\sim}A)$. □

7.2 Semi-paraconsistent adaptive deontic logic

The logic **DCLuN** and its adaptive extensions consistently accommodate all types of normative conflicts that we have encountered so far.

But they also consistently accommodate plain contradictions between formulas not involving deontic operators, such as $p \wedge \sim p$. One could argue that this is overkill. Even if normative conflicts are part of life and should be accommodated in a deontic logic, there is no need to allow also for a non-deontic statement and its negation to be true at the same time.

In this section we mention two ways to adjust **DCLuN** and its adaptive extensions so as to tolerate normative conflicts, without having to tolerate all outright contradictions of the form $A \wedge \sim A$. Casey McGinnis coined the term *semi-paraconsistent deontic logic* for paraconsistent deontic logics that meet this desideratum [McGinnis, 2007b; McGinnis, 2007a].

Excluding non-deontic contradictions The logic $\mathbf{DCLuN_1}$ is obtained by closing **DCLuN** under the axiom schema (Cons_1):[53]

$$\text{Where } A \in \mathcal{W}^{\sim} : \sim A \supset \neg A \qquad (\text{Cons}_1)$$

Where $A \in \mathcal{W}^{\sim}$, (Cons_1) takes care that $A \wedge \sim A$ is trivialized in $\mathbf{DCLuN_1}$. This means that for non-deontic formulas, we obtain full **CL**. Still, $\mathbf{DCLuN_1}$, like **DCLuN**, is highly conflict-tolerant. Where as before $\dagger, \ddagger \in \{\sim, \neg\}$:

$$\mathsf{O}A \wedge \mathsf{O}{\sim}A \not\vdash_{\mathbf{DCLuN_1}} \mathsf{O}B \qquad (57)$$

$$\mathsf{O}A \wedge \mathsf{P}^{\ddagger}_{\dagger}{\sim}A \not\vdash_{\mathbf{DCLuN_1}} \mathsf{O}B \qquad (58)$$

$$\mathsf{O}{\sim}A \wedge \mathsf{P}^{\ddagger}_{\dagger}A \not\vdash_{\mathbf{DCLuN_1}} \mathsf{O}B \qquad (59)$$

$$\mathsf{O}A \wedge {\sim}\mathsf{O}A \not\vdash_{\mathbf{DCLuN_1}} \mathsf{O}B \qquad (60)$$

$$\mathsf{P}^{\ddagger}_{\dagger}A \wedge {\sim}\mathsf{P}^{\ddagger}_{\dagger}A \not\vdash_{\mathbf{DCLuN_1}} \mathsf{O}B \qquad (61)$$

As desired, $\mathbf{DCLuN_1}$ consistently accommodates normative conflicts while trivializing contradictions between statements without occurrences of deontic operators.

Semantically, the logic $\mathbf{DCLuN_1}$ is characterized by imposing the following additional condition on **DCLuN**-models:

$$\text{For all } A \in \mathcal{W}^{\sim} : v(A, w_0) = 1 \text{ iff } v(\sim A, w_0) = 0 \qquad (\text{C}_1^0)$$

Unlike **DCLuN**, the logic $\mathbf{DCLuN_1}$ is not a normal modal logic, since it is not closed under the standard necessitation rule (N). That is, even

[53] Where $\vdash \subseteq \wp(\Phi) \times \Phi$ is a consequence relation and Δ is a set of axioms, we obtain $\vdash_{\Delta}$, *the closure of* $\vdash$ *under* Δ, as follows: $\Gamma \vdash_{\Delta} A$ iff $\Gamma \cup \Delta \vdash A$. This means that one cannot e.g. apply necessitation to members of Δ.

though $\sim p \supset \neg p$ is a theorem of the logic, $\mathsf{O}(\sim p \supset \neg p)$ is not. For similar reasons, the logic is not closed under Uniform Substitution. For instance, $\sim\mathsf{O}p \supset \neg\mathsf{O}p$ is not a theorem of $\mathbf{DCLuN_1}$.

Adaptive logics based on $\mathbf{DCLuN_1}$ can be defined just as before. Mind however that abnormalities of the form $A \wedge \sim A$ for $A \in \mathcal{W}^{\sim}$ are vacuous in the resulting adaptive logics, since they are anyway trivialized by their lower limit logic, in view of (Cons_1). These adaptive logics will perform just as well as $\mathbf{DCLuN^x}$, in that they validate all the inferences from our list of benchmark examples.

Excluding all contradictions at the actual world A second, stronger semi-paraconsistent deontic logic is obtained by closing **DCLuN** under the unrestricted version of (Cons_1):

$$\sim A \supset \neg A \qquad (\text{Cons}_2)$$

Call the resulting logic $\mathbf{DCLuN_2}$. Its semantics is obtained by imposing the following condition on **DCLuN**-models:

$$v(\sim A, w_0) = 1 \text{ iff } v(A, w_0) = 0 \qquad (\text{C}_2^0)$$

In the $\mathbf{DCLuN_2}$-semantics, $\sim$ and $\neg$ are interchangeable at w_0. At all other worlds, $\neg$ remains strictly stronger than $\sim$. This means that contradictions outside the scope of O are trivialized, whereas contradictions within the scope of O are not.

The logic $\mathbf{DCLuN_2}$ is not as conflict-tolerant as $\mathbf{DCLuN_1}$, since it trivializes conflicts of the form $\mathsf{O}A \wedge \sim\mathsf{O}A$ or $\mathsf{P}^{\ddagger}_{\dagger}A \wedge \sim\mathsf{P}^{\ddagger}_{\dagger}A$, where $\dagger, \ddagger \in \{\sim, \neg\}$. Since ($\text{Cons}_2$) and ($\text{C}_2^0$) are no longer restricted to members of $\mathcal{W}^{\sim}$, the logic $\mathbf{DCLuN_2}$ satisfies the rule of uniform substitution, although necessitation (in its full generality) is still invalid.

Just as with **DCLuN** and $\mathbf{DCLuN_1}$, we can use $\mathbf{DCLuN_2}$ as a lower limit logic of our adaptive logic. In this case, the set of abnormalities can be further simplified to the following:

$$\Omega_2^{\sim} = \{\mathsf{P}^{\neg}_{\neg}(A \wedge \sim A) \mid A \in \mathcal{W}^{\sim}\}$$

7.3 Other paraconsistent negations

CLuN is the weakest logic which verifies the full positive fragment of **CL** as well as the principle of Excluded Middle (EM). Stronger paraconsistent logics can be obtained by adding to **CLuN** the double negation

laws and/or de Morgan's laws for negation:

$$\sim\sim A \equiv A \qquad (\text{A}\sim\sim)$$
$$\sim(A \supset B) \equiv (A \wedge \sim B) \qquad (\text{A}\sim\supset)$$
$$\sim(A \wedge B) \equiv (\sim A \vee \sim B) \qquad (\text{A}\sim\wedge)$$
$$\sim(A \vee B) \equiv (\sim A \wedge \sim B) \qquad (\text{A}\sim\vee)$$
$$\sim(A \equiv B) \equiv ((A \vee B) \wedge (\sim A \vee \sim B)) \qquad (\text{A}\sim\equiv)$$

Let **CLuNs** be obtained by adding all of these axioms to **CLuN**. Analogously to the construction of **DCLuN**, we can now construct the logic **DCLuNs** by enriching **CLuNs** with (K), (D), and (N).

One clear difference between **DCLuN**-based ALs and **DCLuNs**-based ALs is that the latter verify a number of additional inferences in a non-defeasible way. For instance, where $\Gamma = \{\mathsf{O}(p \wedge q), \mathsf{O}\sim(p \wedge q)\}$, one cannot **DCLuN$^{\mathbf{r}}$**-derive $\mathsf{O}(\sim p \vee \sim q)$ from Γ, since one cannot rely on the falsehood of the abnormality $\mathsf{P}^{\neg}_{\neg}((p \wedge q) \wedge \sim(p \wedge q))$. In contrast, one can finally **DCLuNs$^{\mathbf{r}}$**-derive $\mathsf{O}(\sim p \vee \sim q)$ from the same premise set, simply in view of properties of **DCLuNs**.

We have to take care when constructing adaptive logics on the basis of **DCLuNs**. Suppose that we work with the set $\Omega^{\sim}$ of **DCLuN$^{\mathbf{x}}$**-abnormalities.

1	$\mathsf{O}p$	Prem	$\emptyset$
2	$\mathsf{O}\sim p$	Prem	$\emptyset$
3	$\mathsf{O}q$	Prem	$\emptyset$
4	$\mathsf{O}(\sim q \vee r)$	Prem	$\emptyset$
5	$\mathsf{O}r$	3,4;RC	$\{\mathsf{P}^{\neg}_{\neg}(q \wedge \sim q)\}\checkmark^{6}$
6	$\mathsf{P}^{\neg}_{\neg}(q \wedge \sim q) \vee \mathsf{P}^{\neg}_{\neg}((p \wedge r) \wedge \sim(p \wedge r))$	1-4;RU	$\emptyset$

Line 5 is marked in view of the minimal Dab-formula derived at line 6. There is no extension of this proof in which to unmark line 5. The proof illustrates that $\mathsf{O}r$ is not finally derivable from the premises at lines 1-4. This is counter-intuitive.

If we are to build an adaptive logic on the basis of the lower limit logic **DCLuNs** and the set of abnormalities $\Omega^{\sim}$, the resulting logic would exhibit flip-flop behavior (see Section 5 where we also encountered this problem). The solution is to restrict the set of abnormalities as follows:

$$\Omega^{\sim}_{s} = \{A \wedge \sim A \mid A \in \mathcal{S}\} \cup \{\mathsf{O}A \wedge \sim \mathsf{O}A \mid A \in \mathcal{W}^{\sim}\} \cup \{\mathsf{P}^{\neg}_{\neg}(A \wedge \sim A) \mid A \in \mathcal{S}\} \qquad (62)$$

Given (A$\sim\sim$)-(A$\sim\equiv$), inconsistencies between complex formulas in $\mathcal{W}$ can be reduced to inconsistencies at the level of atoms in **DCLuNs**. In view of this, **DCLuNs$^{\mathsf{x}}$**-abnormalities must be restricted accordingly, on pain of flip-flop behavior. That is, where $A \in \mathcal{W}$, $A \wedge \sim A$ and $\mathsf{P}^{\neg}(A \wedge \sim A)$ only counts as an abnormality when $A \in \mathcal{S}$.

The situation is different for formulas of the form $\mathsf{O}A \wedge \sim\mathsf{O}A$: within the scope of O, inconsistencies between complex formulas do *not* reduce to inconsistencies at the level of atoms. For instance, the inference from $\mathsf{O}(p \wedge q) \wedge \sim\mathsf{O}(p \wedge q)$ to $(\mathsf{O}p \wedge \sim\mathsf{O}p) \vee (\mathsf{O}q \wedge \sim\mathsf{O}q)$ is not **DCLuNs**-valid, since $\sim\mathsf{O}(p \wedge q)$ does not **DCLuNs**-entail $\sim\mathsf{O}p \vee \sim\mathsf{O}q$. More generally, where A is a complex formula, the formula $\sim\mathsf{O}A$ cannot be further analysed in **DCLuNs**. So, as in **DCLuN$^{\mathsf{x}}$**, all formulas of the form $\mathsf{O}A \wedge \sim\mathsf{O}A$ count as abnormalities in **DCLuNs$^{\mathsf{x}}$**.

Let **DCLuNs$^{\mathsf{x}}$** be the adaptive logic defined by the lower limit logic **DCLuNs**, the set of abnormalities $\Omega_s^{\sim}$, and the strategy $\mathsf{x} \in \{\mathsf{r}, \mathsf{m}\}$. Then clearly the formula derived at line 6 of the proof above is no longer a minimal Dab-formula, and line 5 remains unmarked. We can still derive the Dab-formula $\mathsf{P}^{\neg}(q \wedge \sim q) \vee \mathsf{P}^{\neg}(p \wedge \sim p) \vee \mathsf{P}^{\neg}(r \wedge \sim r)$ from lines 1-4 via RU, in view of

$$\mathsf{P}^{\neg}((p \wedge r) \wedge \sim(p \wedge r)) \vdash_{\mathbf{DCLuNs}} \mathsf{P}^{\neg}(p \wedge \sim p) \vee \mathsf{P}^{\neg}(r \wedge \sim r) \qquad (63)$$

However, this Dab-formula is not minimal, since its disjunct $\mathsf{P}^{\neg}(p \wedge \sim p)$ is a **DCLuNs**-consequence of the formulas $\mathsf{O}p$ and $\mathsf{O}\sim p$ at lines 1 and 2. As a result, line 5 is finally derivable and $\mathsf{O}r$ is a **DCLuNs$^{\mathsf{x}}$**-consequence of the premises.

Other than **CLuN** and **CLuNs**, there is a wide variety of paraconsistent logics that can serve as the core logic of an inconsistency-adaptive logic. We could, for instance, treat '$\sim$' as a dummy operator for which not even (EM) holds by removing (A$\sim$1) in the axiomatization of **CLuN**. The resulting logic is called **CLoN** (for Classical Logic with both gluts and gaps for Negation). Extending **CLoN** with (A$\sim\sim$)-(A$\sim\equiv$) results in the logic **CLoNs**. These systems too can be extended deontically and adaptively. In addition, one can also consider semi-paraconsistent versions of **DCLuNs** and **DCLoNs**.

7.4 Further reading and open ends

For a general overview of paraconsistent logic, see e.g. [Priest, 2002; Priest *et al.*, 2015]. For an overview of (monotonic) paraconsistent deontic logic, we refer to [Goble, 2013, Sect. 6.1] in volume 1 of this handbook.

The first paper on inconsistency-adaptive logic – published in 1989, but written in 1981 – is [Batens, 1989], where the proof theory for the reliability strategy was first presented. The minimal abnormality strategy was first presented (semantically) in [Batens, 1986]. The (propositional) results of the two aforementioned papers were generalized to the predicative level in [Batens, 1999a]. For an overview and more recent results within the inconsistency-adaptive program, see [Batens, 2015].

Inconsistency-adaptive deontic logics were presented in [Beirlaen, 2012; Beirlaen *et al.*, 2013], in [Beirlaen and Straßer, 2011], and in [Goble, 2014]. Most of these systems – in contrast to the ones presented in this section – allow for the following inference:[54]

$$\mathsf{O}A \wedge \mathsf{O}{\sim}A \vdash {\sim}\mathsf{O}{\sim}A \wedge \mathsf{O}{\sim}A \tag{64}$$

That is, conflicts of the form $\mathsf{O}A \wedge \mathsf{O}{\sim}A$ entail plain contradictions. Goble is critical of such systems:

> That seems an exceedingly strong commitment. It is easy to accept that there are normative conflicts, harder to suppose they all yield contradictions that are true. Even Priest, the hierarch of dialetheism, does not consider normative conflicts so paradoxical [Goble, 2014, Fn. 15].

The systems presented in this section circumvent Goble's criticism by invalidating inferences like (64).

In [Beirlaen and Straßer, 2013] the semi-paraconsistent deontic logic **LNP** is presented and extended within the adaptive logics framework. **LNP** is a close cousin of $\mathbf{DCLoNs_2}$, but has a slightly different language in which the P-operator is primitive, and in which '$\neg$' is allowed only outside the scope of deontic operators, while '$\sim$' is allowed only inside the scope of deontic operators.

Once we are open to the possibility of changing the logic of the connectives, new questions arise. For instance, why should we always blame negation for the explosive behavior of a logic, and why not weaken the meaning of the other connectives? Why not e.g. give up addition for $\vee$ (i.e., to derive $A \vee B$ from A or from B)? In [Batens, 1999b], Batens shows that a whole range of interesting new logics come to the fore, once we generalize the idea of gluts and gaps to other connectives and logical

[54] (64) holds for the inconsistency-adaptive systems presented in [Beirlaen, 2012; Beirlaen *et al.*, 2013], and [Beirlaen and Straßer, 2011]. The closely related principle $\mathsf{O}A \wedge \mathsf{O}{\sim}A \vdash (\mathsf{O}{\sim}A \wedge {\sim}\mathsf{O}{\sim}A) \vee \mathsf{O}B$ holds for those logics mentioned in [Goble, 2014] which satisfy the 'deontic addition' schema $\mathsf{O}A \supset \mathsf{O}(A \vee B)$.

operators. The application of all this to deontic reasoning is yet to be studied in detail, but it can draw on many existing results concerning corrective ALs.

In [Beirlaen and Straßer, 2014], a very rich paraconsistent deontic logic is presented, one that allows the user to express not only obligations that concern states of affairs, but also obligations that concern agency. The language of these systems contains modal operators $\Box_J$ for "the group of agents J brings it about that", inspired by existing work on logics of agency [Segerberg, 1992; Belnap and Perloff, 1993; Elgesem, 1997]. This in turn allows one to distinguish between various different types of *inter-personal* and *intra-personal* deontic conflicts:[55]

$$\mathsf{O}\Box_i A \wedge \mathsf{O}\Box_j \sim A \tag{65}$$
$$\mathsf{O}\Box_i A \wedge \mathsf{P}\Box_j \sim A \tag{66}$$
$$\mathsf{O}\Box_i A \wedge \mathsf{O}\Box_i \sim A \tag{67}$$
$$\mathsf{O}\Box_i A \wedge \mathsf{P}\Box_i \sim A \tag{68}$$
$$\mathsf{O}\Box_i A \wedge \mathsf{O}\sim\Box_i A \tag{69}$$
$$\mathsf{O}\Box_i A \wedge \mathsf{P}\sim\Box_i A \tag{70}$$
$$\mathsf{O}\Box_i A \wedge \sim\mathsf{O}\Box_i A \tag{71}$$
$$\mathsf{P}\Box_i A \wedge \sim\mathsf{P}\Box_i A \tag{72}$$

One further advantage of such richer formal languages in the context of adaptive reasoning is that they allow us to prioritize the minimization of certain types of conflicts over that of others. For instance, we may consider conflicts of type (67) worse than those of type (65) and (66), since the former clearly violate the principle that if an agent ought to bring about A, then that agent is also able to see to A – assuming agents cannot bring about contradictions. Such a prioritized reasoning can be modeled in terms of a lexicographic AL (cf. Section 3.4).

8 Conflict-tolerant adaptive logics: round-up

In this section, we give an overview of the main features of the logics discussed so far. We start by giving an overview of the performance of

[55] An inter-personal conflict is one that holds between the obligations of different agents, whereas an intra-personal conflict obtains between the obligations of a single agent. One famous example of an inter-personal normative conflict can be found in Sophocles' *Antigone*, where due to the city's laws, Creon is obliged to prevent the burial or Antigon's brother Polyneices, but Antigone faces a religious and familial obligation to bury Polyneices [Marcus, 1980; Gowans, 1987].

revisionist ALs with respect to the criteria introduced in Section 4.2. In Section 8.2 we return to the logics from Section 3. We show how these can be evaluated using similar criteria, and how they can be enriched in various ways.

8.1 Revisionist deontic adaptive logics: overview

The behavior of the revisionist adaptive logics with respect to the criteria from Section 4.2 is summarized in Tables 1 and 2. Principles (arguments) that are valid in a given logic receive a ✓, invalid principles (arguments) receive a ✗.[56] Where the premises of an argument are trivialized by a given logic, we write a $\bot$ in Table 2.

	DEX	DEX-O$\bot$	DEX-P$\bot$	DEX-OP$\neg$	DEX-O$\neg$P
LUM.a$^{\mathbf{x}}$	✗	✗	✓	✓	✗
LUM.b$^{\mathbf{x}}$	✗	✓	✓	✓	✗
LUM.c$^{\mathbf{x}}$	✗	✓	✓	✓	✗
P$^{\mathbf{x}}$	✗	✓	✓	✓	✗
DCLuN$^{\mathbf{x}}$	✗	✗	✗	✗	✗
DCLuN$^{\mathbf{x}}_{\mathbf{1}}$	✗	✗	✗	✗	✗
DCLuN$^{\mathbf{x}}_{\mathbf{2}}$	✗	✗	✗	✗	✓

Table 1: Behavior of deontic ALs with respect to various explosion principles.

It should be noted here once more (in line with our remarks in Section 4.2) that whether a given AL validates some form of deontic explosion or a specific inference should not be seen as conclusive evidence in favour of or against such a logic. The above tables are mostly for purposes of comparison and classification, and do not serve as strict criteria of the relative success or failure of the respective systems or their purposes. For example, with a view to supporting ought-implies-can, a system might be *designed* to consider $\mathsf{O}(A \wedge \neg A)$ inconsistent even while $\mathsf{O}A \wedge \mathsf{O}\neg A$ is consistent. In that case, that the system validates (DEX-O$\bot$) may be taken as a virtue rather than a vice. Likewise, the validation of (DEX-

[56] As noted before, for the logics from Section 7 we assume that the principles (arguments) in question are formalized using the paraconsistent negation sign $\sim$.

	S	J	R1	R2	T	N1	N2
LUM.a$^{\mathbf{r}}$	✓	✓	✗	✗	✗	⊥	✗
LUM.a$^{\mathbf{m}}$	✓	✓	✗	✗	✗	⊥	✓
LUM.b$^{\mathbf{r}}$	✓	✓	✗	✗	✗	⊥	✗
LUM.b$^{\mathbf{m}}$	✓	✓	✗	✗	✗	⊥	✓
LUM.c$^{\mathbf{r}}$	✓	✓	✗	✗	✗	✗	✗
LUM.c$^{\mathbf{m}}$	✓	✓	✗	✗	✗	✓	✓
P$^{\mathbf{x}}$	✓	✓	✓	✓	✓	✓	✓
DCLuN$^{\mathbf{x}}$	✓	✓	✓	✓	✓	✓	✓
DCLuN$_{\mathbf{1}}^{\mathbf{x}}$	✓	✓	✓	✓	✓	✓	✓
DCLuN$_{\mathbf{2}}^{\mathbf{x}}$	✓	✓	✓	✓	✓	⊥	✓
DCLuNs$^{\mathbf{x}}$	✓	✓	✓	✓	✓	✓	✓

Table 2: Behavior of deontic ALs with respect to the Smith (S), Jones (J), Roberts (R1 and R2), Thomas (T), and Natascha (N1 and N2) arguments from Section 4.)

$\mathsf{OP}\neg$) would be embraced by one with a classical point of view (and given the standard interdefinability of O and P).

Let us close this overview with a technical point. All ALs discussed in Sections 5–7 have a monotonic, conflict-tolerant deontic logic as their lower limit logic. The latter logics are mutually incomparable, in the sense that none is stronger than any other.[57] For instance, the logic **LUM.a** from Section 5 invalidates (Inh) but validates (Agg); conversely, the logic **P** that is discussed in Section 6 invalidates (Agg) but validates (Inh). It can easily be shown that any two ALs that are based on such incomparable lower limit logics, are themselves equally incomparable. This is an immediate corollary of the following:[58]

[57]A small warning is in place here. The paraconsistent deontic logics of the **DCLuN**-family, presented in Section 7, work with a richer language that contains both a paraconsistent and a classical negation. The claim we make here concerns the fragment of those logics *without* the classical negation.

[58]Theorem 8.1 generalizes one direction of Theorem 3.3 in [Van De Putte and Straßer, 2014].

Theorem 8.1. *Let* $\mathbf{AL_1}$ *and* $\mathbf{AL_2}$ *be two ALs in standard format, defined by the triples* $\langle \mathbf{LLL_1}, \Omega_1, \mathsf{x}_1 \rangle$*, resp.* $\langle \mathbf{LLL_2}, \Omega_2, \mathsf{x}_2 \rangle$*, over a given formal language. If* $\vdash_{\mathbf{AL_1}} \subseteq \vdash_{\mathbf{AL_2}}$*, then* $\vdash_{\mathbf{LLL_1}} \subseteq \vdash_{\mathbf{LLL_2}}$*.*

Proof. By contraposition: suppose that $\vdash_{\mathbf{LLL_1}} \not\subseteq \vdash_{\mathbf{LLL_2}}$. Let Γ, A be such that (i) $\Gamma \vdash_{\mathbf{LLL_1}} A$ but (ii) $\Gamma \nvdash_{\mathbf{LLL_2}} A$. By (i) and the monotonicity of $\mathbf{LLL_1}$, (iii) $\Gamma \cup \{\neg A\} \vdash_{\mathbf{LLL_1}} A$. by (ii), $\Gamma \cup \{\neg A\}$ is $\mathbf{LLL_2}$-consistent, and hence by **CL**-properties, (iv) $\Gamma \cup \{\neg A\} \nvdash_{\mathbf{LLL_2}} A$. By (iii) and Theorem 3.15, $\Gamma \cup \{\neg A\} \vdash_{\mathbf{AL_1}} A$. By (iv) and Theorem 3.11, $\Gamma \cup \{\neg A\} \nvdash_{\mathbf{AL_2}} A$. Hence, $\vdash_{\mathbf{AL_1}} \not\subseteq \vdash_{\mathbf{AL_2}}$. □

As a result, the ALs discussed in Sections 5-7 are incomparable, i.e. an AL belonging to one of these three types cannot in general be stronger or weaker than an AL belonging to another of the three types.

8.2 *Prima facie* obligations revisited

Explosion principles To apply the criteria from Section 4.2 to the logics from Section 3.1, we need some more preparation. We take it that the premises of the explosion principles, resp. arguments under consideration are all concerned with *prima facie* obligations, whereas their conclusion concerns actual obligations. Under this translation, $\mathbf{SDL^r_p}$ and $\mathbf{SDL^m_p}$ invalidate the analogues of (DEX) and (DEX-O⊥):

$$\mathsf{O^p}A \wedge \mathsf{O^p}\neg A \vdash \mathsf{O}B \tag{73}$$

$$\mathsf{O^p}(A \wedge \neg A) \vdash \mathsf{O}B \tag{74}$$

The other explosion principles cannot as easily be translated to these systems, because in Section 3.1 we did not define a corresponding *prima facie* permission operator for the logics $\mathbf{SDL^x_p}$.

Suppose that we add a second dummy operator $\mathsf{P^p}$ to the language of $\mathbf{SDL_p}$. For the adaptive extension of the resulting logic, we re-define the set of abnormalities Ω_p by including both formulas of the form $\mathsf{O^p}A \wedge \neg\mathsf{O}A$ *and* formulas of the form $\mathsf{P^p}A \wedge \neg\mathsf{P}A$. In the resulting logic, the following analogues of the explosion principles (DEX-P⊥) and (DEX-OP¬) are invalid:

$$\mathsf{P^p}(A \wedge \neg A) \vdash \mathsf{O}B \tag{75}$$

$$\mathsf{O^p}A \wedge \mathsf{P^p}\neg A \vdash B \tag{76}$$

$$\mathsf{O^p}A \wedge \neg\mathsf{P^p}A \vdash B \tag{77}$$

Note that conflicts of the form $\mathsf{O}^{\mathsf{p}}A \wedge \mathsf{P}^{\mathsf{p}}\neg A$ give rise to disjunctions of abnormalities in this logic:

$$\mathsf{O}^{\mathsf{p}}A \wedge \mathsf{P}^{\mathsf{p}}\neg A \vdash (\mathsf{O}^{\mathsf{p}}A \wedge \neg\mathsf{O}A) \vee (\mathsf{P}^{\mathsf{p}}\neg A \wedge \neg\mathsf{P}\neg A) \tag{78}$$

In case there is a conflict between a *prima facie* obligation and a *prima facie* permission, the adaptive logic will not prioritize one over the other. This is in line with [Hansen, 2014], where it is argued that permission should not take priority over obligations or conversely. Should one nevertheless want a logic that does treat one type of conflict as "worse" than the other, then one can turn to the format of lexicographic ALs as sketched in Section 3.4.

Benchmark examples First, in both $\mathbf{SDL}^{\mathbf{r}}_{\mathbf{p}}$ and $\mathbf{SDL}^{\mathbf{m}}_{\mathbf{p}}$, the Smith and Jones arguments are $\mathbf{SDL}^{\mathbf{x}}_{\mathbf{p}}$-valid, while Roberts and Thomas are not.

$$\mathsf{O}^{\mathsf{p}}(f \vee s), \mathsf{O}^{\mathsf{p}}\neg f \vdash_{\mathbf{SDL}^{\mathbf{x}}_{\mathbf{p}}} \mathsf{O}s \qquad \text{(Smith)}$$

$$\mathsf{O}^{\mathsf{p}}(j \wedge s) \vdash_{\mathbf{SDL}^{\mathbf{x}}_{\mathbf{p}}} \mathsf{O}j \qquad \text{(Jones)}$$

$$\mathsf{O}^{\mathsf{p}}(t \wedge r), \mathsf{O}^{\mathsf{p}}(\neg t \wedge v) \nvdash_{\mathbf{SDL}^{\mathbf{x}}_{\mathbf{p}}} \mathsf{O}r \wedge \mathsf{O}v \qquad \text{(Roberts 1)}$$

$$\mathsf{O}^{\mathsf{p}}(t \wedge r), \mathsf{O}^{\mathsf{p}}(\neg t \wedge v) \nvdash_{\mathbf{SDL}^{\mathbf{x}}_{\mathbf{p}}} \mathsf{O}(r \wedge v) \qquad \text{(Roberts 2)}$$

$$\mathsf{O}^{\mathsf{p}}(t \wedge (f \vee s)), \mathsf{O}^{\mathsf{p}}(\neg t \wedge \neg f) \nvdash_{\mathbf{SDL}^{\mathbf{x}}_{\mathbf{p}}} \mathsf{O}s \qquad \text{(Thomas)}$$

In order to infer the conclusions of the Roberts and Thomas arguments, we would need to detach the obligations $\mathsf{O}(t \wedge r)$ and $\mathsf{O}(t \wedge (f \vee s))$ respectively. But we cannot do that in view of the following minimal Dab-consequences of the respective premise sets:

$$(\mathsf{O}^{\mathsf{p}}(t \wedge r) \wedge \neg\mathsf{O}(t \wedge r)) \vee (\mathsf{O}^{\mathsf{p}}(\neg t \wedge v) \wedge \neg\mathsf{O}(\neg t \wedge v)) \tag{79}$$

$$(\mathsf{O}^{\mathsf{p}}(t \wedge (f \vee s)) \wedge \neg\mathsf{O}(t \wedge (f \vee s))) \vee (\mathsf{O}^{\mathsf{p}}(\neg t \wedge \neg f) \wedge \neg\mathsf{O}(\neg t \wedge \neg f)) \tag{80}$$

One way of accounting for the Roberts and Thomas arguments is to strengthen $\mathbf{SDL}^{\mathbf{x}}_{\mathbf{p}}$ by closing the operator O^{p} under a number of further rules. For instance, we could add a principle permitting the inference from $\mathsf{O}^{\mathsf{p}}(A \wedge B)$ to $\mathsf{O}^{\mathsf{p}}A$, such as (Inh). That would enable us to infer $\mathsf{O}^{\mathsf{p}}r$ given $\mathsf{O}^{\mathsf{p}}(t \wedge r)$, and $\mathsf{O}r$ given $\mathsf{O}^{\mathsf{p}}r$ (on the condition $\mathsf{O}^{\mathsf{p}}r \wedge \neg\mathsf{O}r$). Clearly, however, not anything goes when closing O^{p} under additional rules. For one thing, we do not want to end up with full **SDL** or even **K** for *prima facie* obligations, as this would completely annihilate our

initial objective. But also if we characterize O^{P} in terms of weaker logics like the ones presented in Sections 5-7, we should be careful. After all, the richer one's lower limit logic, the more likely one is to end up with flip-flop problems that will require further tinkering with the set of abnormalities, much as we had to do in previous sections.

For the Natascha argument, one can translate the impossibility of $s \wedge m$ using the operator O for actual obligations. The underlying idea is that constraints concerning what is practically (im)possible only have a bearing on actual obligations, not on the *prima facie* obligations. This can again be done in two different ways, giving rise to two different premise sets. For both, the validity of the argument will depend on the adaptive strategy:

$$\mathsf{O}^{\mathsf{P}}s, \mathsf{O}^{\mathsf{P}}m, \mathsf{O}^{\mathsf{P}}(s \supset t), \mathsf{O}^{\mathsf{P}}(m \supset t), \neg\mathsf{O}(s \wedge m) \nvdash_{\mathbf{SDL}^{\mathbf{r}}_{\mathbf{p}}} \mathsf{O}t \quad \text{(Natascha 1)}$$

$$\mathsf{O}^{\mathsf{P}}s, \mathsf{O}^{\mathsf{P}}m, \mathsf{O}^{\mathsf{P}}(s \supset t), \mathsf{O}^{\mathsf{P}}(m \supset t), \neg\mathsf{O}(s \wedge m) \vdash_{\mathbf{SDL}^{\mathbf{m}}_{\mathbf{p}}} \mathsf{O}t \quad \text{(Natascha 1)}$$

$$\mathsf{O}^{\mathsf{P}}s, \mathsf{O}^{\mathsf{P}}m, \mathsf{O}^{\mathsf{P}}(s \supset t), \mathsf{O}^{\mathsf{P}}(m \supset t), \mathsf{O}\neg(s \wedge m) \nvdash_{\mathbf{SDL}^{\mathbf{r}}_{\mathbf{p}}} \mathsf{O}t \quad \text{(Natascha 2)}$$

$$\mathsf{O}^{\mathsf{P}}s, \mathsf{O}^{\mathsf{P}}m, \mathsf{O}^{\mathsf{P}}(s \supset t), \mathsf{O}^{\mathsf{P}}(m \supset t), \mathsf{O}\neg(s \wedge m) \vdash_{\mathbf{SDL}^{\mathbf{m}}_{\mathbf{p}}} \mathsf{O}t \quad \text{(Natascha 2)}$$

In Sections 4–8 we defined and discussed a large variety of conflict-tolerant deontic logics that can be developed within the AL framework. More variation is possible, as there are other ways still to define conflict-tolerant deontic logics – by moving to a hyperintensional framework, for instance – and strengthen them adaptively. Moreover, existing systems can be altered by making them more expressive, e.g. by considering the interplay between deontic modalities and alethic, doxastic, or epistemic modalities. All this goes to show that adaptive logics provide a versatile and modular framework for conflict-tolerant normative reasoning, and that their applications to this problem are far from exhausted.

9 Conditional obligations and adaptive detachment

SDL is inadequate not just for accommodating normative conflicts in deontic logic, but also for representing deontic conditionals, as we will explain below.[59] Within the vast literature on such conditionals, one

[59] We will only sketch the latter inadequacy here. It is discussed at length in Section 8.5. and in the Appendix of Ch. 1 in the first volume of this handbook. For other overviews of this problem, see for instance [Åqvist, 2002; Carmo and Jones, 2002].

can distinguish three general approaches. The first is to represent them by means of a *dyadic* obligation operator $\mathsf{O}(\cdot \mid \cdot)$, and to read a formula $\mathsf{O}(B \mid A)$ as 'If A, then B is obligatory'. A second approach is to treat the problems surrounding deontic conditionals as symptomatic of the bigger challenge of how to formalize conditional statements in general. The third approach is more abstract: it treats deontic conditionals as pairs connecting a given "input" with an "output", and defines specific proof theories and an operational semantics (based on the principle of detachment and **CL**) for such connections.

We will discuss these three different approaches in Sections 9.1-9.3 respectively, showing how the framework of ALs can be useful in each of them. Our discussion will be mainly tentative; we provide pointers to more technical results and fully worked-out proposals in the literature at the end of each subsection.

9.1 Adaptive dyadic deontic logics

Helping one's neighbours Let us illustrate the distinctive problems surrounding deontic conditionals by means of a so-called Chisholm scenario – after [Chisholm, 1963]. This scenario can be represented as follows in the dyadic setting:

(i) It is obligatory that Jones goes to the aid of his neighbours ($\mathsf{O}g$).

(ii) It is obligatory that if Jones goes to the aid of his neighbours, then he tells them he is coming ($\mathsf{O}(t \mid g)$).

(iii) If Jones does not go to the aid of his neighbours, then he ought not to tell them he is coming ($\mathsf{O}(\neg t \mid \neg g)$).

(iv) Jones does not go to the aid of his neighbours ($\neg g$).

Recall now the principles of *factual detachment* (FD) and *deontic detachment* (DD) from Section 1:

$$A, \mathsf{O}(B \mid A) \vdash \mathsf{O}B \qquad \text{(FD)}$$
$$\mathsf{O}A, \mathsf{O}(B \mid A) \vdash \mathsf{O}B \qquad \text{(DD)}$$

Given premises (iii) and (iv), we can use (FD) to infer an obligation $\mathsf{O}\neg t$ for Jones not to tell his neighbours he is coming. However, given premises (i) and (ii), we can use (DD) to infer an obligation $\mathsf{O}t$ for Jones to tell his neighbours he is coming.

But now we face a dilemma. Jones cannot both tell and not tell his neighbours he is coming. So, each of (DD) and (FD) has some intuitive appeal, but together they lead to a deontic conflict, and hence explosion if the logic of O is **SDL**. This is the dilemma of deontic and factual detachment, also known in the literature as "the dilemma of detachment and commitment" [Åqvist, 2002; van Eck, 1982]. In fact, one should rather speak here of a *trilemma*, since one may deny that **SDL** is an appropriate logic for obligations, and insist that both (FD) and (DD) should be unconditionally valid. This means one needs a conflict-tolerant deontic logic for O, much as those discussed in preceding sections. Here, we will first focus on the other two horns of the trilemma and exclude conflicts at the level of O.

Since each of (DD) and (FD) seems reasonable in isolation, Hilpinen and McNamara argue that we cannot just pick one of them at the expense of the other, and that we need to move to a more nuanced position beyond this choice [Hilpinen and McNamara, 2013, p. 119]. One solution is to make the detachment – via (DD) or (FD) – of unconditional obligations subject to further conditions, such as joint consistency. The AL framework allows us to make this idea exact, and to study its pros and cons.

A simple solution Let $\mathbf{SDL_d}$ be the logic obtained by replacing the unary *prima facie* operator $\mathsf{O}^{\mathsf{P}}(\cdot)$ of $\mathbf{SDL_p}$ with the conditional operator $\mathsf{O}(\cdot \mid \cdot)$. As we did with the O^{P}-operator of $\mathbf{SDL_p}$, we treat the new conditional operator like a dummy operator in $\mathbf{SDL_d}$.

Some authors treat unconditional obligations $\mathsf{O}A$ on the same foot as conditional obligations of the type $\mathsf{O}(A \mid \top)$. Note that in $\mathbf{SDL_d}$ these are not equivalent. For instance, the conjunction $\mathsf{O}(A \mid \top) \wedge \mathsf{O}(\neg A \mid \top)$ is $\mathbf{SDL_d}$-consistent, while the conjunction $\mathsf{O}A \wedge \mathsf{O}\neg A$ is not. In line with the interpretation in Section 3, $\mathsf{O}(A \mid \top)$ expresses something like "A is an unconditional prima facie obligation", whereas the intended reading of $\mathsf{O}A$ is that "A is an actual obligation".

In order to detach unconditional obligations from conditional obligations, we strengthen $\mathbf{SDL_d}$ adaptively to the logics $\mathbf{SDL_d^x}$, which are defined by the triple $\langle \mathbf{SDL_d}, \Omega_d, \mathsf{x} \rangle$, with $\mathsf{x} \in \{\mathsf{r}, \mathsf{m}\}$ and $\Omega_d = \Omega_{fd} \cup \Omega_{dd}$:

$$\Omega_{fd} = \{\mathsf{O}(B \mid A) \wedge A \wedge \neg \mathsf{O}B \mid A, B \in \mathcal{W}\}$$
$$\Omega_{dd} = \{\mathsf{O}(B \mid A) \wedge \mathsf{O}A \wedge \neg \mathsf{O}B \mid A, B \in \mathcal{W}\}$$

In view of the $\mathbf{SDL_d}$-valid inferences (81) and (82), the adaptive

logics $\mathbf{SDL^x_d}$ allow for the conditional application of (FD) and (DD):

$$A, \mathsf{O}(B \mid A) \vdash \mathsf{O}B \vee (\mathsf{O}(B \mid A) \wedge A \wedge \neg \mathsf{O}B) \tag{81}$$

$$\mathsf{O}A, \mathsf{O}(B \mid A) \vdash \mathsf{O}B \vee (\mathsf{O}(B \mid A) \wedge \mathsf{O}A \wedge \neg \mathsf{O}B) \tag{82}$$

We illustrate the resulting logic by applying it to the Chisholm scenario in (i)-(iv):

1	$\mathsf{O}g$	Prem	$\emptyset$
2	$\mathsf{O}(t \mid g)$	Prem	$\emptyset$
3	$\mathsf{O}(\neg t \mid \neg g)$	Prem	$\emptyset$
4	$\neg g$	Prem	$\emptyset$
5	$\mathsf{O}t$	1,2; RC	$\{\mathsf{O}(t \mid g) \wedge \mathsf{O}g \wedge \neg \mathsf{O}t\}\checkmark^7$
6	$\mathsf{O}\neg t$	3,4; RC	$\{\mathsf{O}(\neg t \mid \neg g) \wedge \neg g \wedge \neg \mathsf{O}\neg t\}\checkmark^7$
7	$(\mathsf{O}(t \mid g) \wedge \mathsf{O}g \wedge \neg \mathsf{O}t) \vee (\mathsf{O}(\neg t \mid \neg g) \wedge \neg g \wedge \neg \mathsf{O}\neg t)$	1-4; RU	$\emptyset$

Lines 4 and 5 remain marked in any extension of this proof, so that neither $\mathsf{O}t$ nor $\mathsf{O}\neg t$ is an $\mathbf{SDL^x_d}$-consequence of the premises at lines 1-4. Thus, in cases of conflict, the applications of (FD) and (DD) that lead to the conflict are rejected.

Some have taken a bolder stance here by arguing that when factual and deontic detachment lead to a conflict, (FD) overrules (DD) or vice versa. We will not go into this discussion here – see [Hilpinen and McNamara, 2013, p. 112-124] for an overview of the various positions. However, let us briefly indicate how this idea of overruling can be modeled with the AL framework.

Recall the lexicographic ALs that were introduced in Section 3.4. Consider the lexicographic ALs defined in terms of the lower limit logic $\mathbf{SDL_d}$ and the sequence $\langle \Omega_{fd}, \Omega_{dd} \rangle$. The idea is that we treat abnormalities with respect to factual detachment as "worst", and hence give priority to (FD) over (DD). For instance, in the Chisholm case, the abnormality $\mathsf{O}(\neg t \mid \neg g) \wedge \neg g \wedge \neg \mathsf{O}\neg t$ will be avoided, and hence the abnormality $\mathsf{O}(t \mid g) \wedge \mathsf{O}g \wedge \neg \mathsf{O}t$ will be assumed to hold. Thus, in such logics, one can conclude that Jones ought not to tell his neighbours he is coming. Other applications of (DD) that do not result in conflicting obligations will remain valid in such logics. Finally, if two different applications of (FD) conflict, they will both be blocked in the adaptive logics.

A (prioritized) combination of various sorts of adaptive reasoning may also be useful for those who insist on the intuitiveness of (FD) and

(DD), and use these to cast doubt on the validity of full **SDL** for O (cf. our discussion of the trilemma of detachment and commitment, supra). Here, one may combine insights and techniques from Sections 5–7 with those from the present section, treating each of (FD), (DD), and (some or all) rules and axioms of **SDL** as defeasible. This way one cannot only accommodate deontic conflicts that arise from an applications of either (FD) or (DD) or both – by invalidating those applications – but also conflicting obligations that happen to be simply there, "unconditionally". In such a setting, one may e.g. prioritize the standard behavior of O over the applicability of (FD) and (DD), thus capturing the intuition that even if they are sometimes to be accepted, deontic conflicts should be avoided whenever possible.

Open problems and further reading The first monotonic dyadic deontic logics were introduced in Bengt Hansson's seminal paper [Hansson, 1969]. See the first chapter of this handbook volume for a detailed study of the history and metatheory of those logics, see Parent's Chapter 1 in this Volume. Hansson-style dyadic deontic logics typically invalidate (FD), while some of them validate (DD).

More recently, van Benthem, Grossi and Liu have investigated the relation between modal logics of preferences, priority structures, and dyadic deontic logic more generally [van Benthem *et al.*, 2014]. In this account, the factual information in the antecedent of (FD) is formalized as a dynamic epistemic event, rather than as a "mere" factual (propositional) statement. This way, the non-monotonicity of reasoning with dyadic obligations is formalized at the object-level, rather than as a property of the consequence relation.

Our focus in this section was on the defeasible application of the detachment principles (FD) and (DD), in a language with both a dyadic operator $\mathsf{O}(\cdot \mid \cdot)$ for conditional obligations and an independent, monadic operator O that satisfies full **SDL**. We did not discuss other logical properties of $\mathsf{O}(\cdot \mid \cdot)$, and instead treated it as a dummy operator much like we treated the O^{p}-operator from Section 3. But we may of course wonder whether there are no logical properties which the dyadic operator ought to satisfy unrestrictedly. Possible candidates include, for instance, the dyadic versions of the aggregation and inheritance principles:

$$(\mathsf{O}(B \mid A) \wedge \mathsf{O}(C \mid A)) \supset (\mathsf{O}(B \wedge C \mid A)) \qquad \text{(DAgg)}$$

$$\text{From } \mathsf{O}(B \mid A) \text{ and } \vdash B \supset C, \text{ to infer } \mathsf{O}(C \mid A) \qquad \text{(DInh)}$$

However, one has to be careful again, since enriching one's lower limit logic may easily give rise to flip-flop-problems, analogous to the monadic

deontic logics presented in previous sections. The solutions that were discussed in those sections may in turn be transferred to the dyadic setting.

Different preferences regarding the characterization of $\mathsf{O}(\cdot \mid \cdot)$ have given rise to a wide variety of dyadic systems, including a range of conflict-tolerant dyadic systems which could in turn be extended adaptively so as to gain further inferential power. For instance, in [Straßer, 2010] and [Straßer, 2014, Ch. 11], Christian Straßer studied conditional versions of some of the **LUM**-systems from Section 5, and presented a number of adaptive extensions of these logics. In [Straßer, 2011] and [Straßer, 2014, Chapters 11–12], Straßer presents a general method for turning dyadic deontic logics into ALs which allow for the conditional application of (FD), paying special attention to Chisholm-scenarios.

Finally, it should also be noted that, even if we leave (FD) and (DD) aside, all the observations and techniques from Sections 5–7 could be applied just as well to the case of dyadic deontic logics as developed, building on Goble's work in [Goble, 2003; Goble, 2004b]. Here again, we may use adaptive logics to steer a middle course between all-too-weak conflict tolerant dyadic systems and deontic explosion.

9.2 Adaptive reasoning with conditionals

Adaptive detachment, generalized Instead of using a binary operator for conditional obligation, one may also introduce a new conditional $\Rightarrow$, so that the logic of deontic conditionals derives from the logic for this new conditional and the logic for the monadic operator O of one's choice. In this section we focus on this second approach.

Suppose we formalize "If A, then B is obligatory" as $A \Rightarrow \mathsf{O}B$.[60] Then at the very least we want to be able to factually detach $\mathsf{O}B$ given A and $A \Rightarrow \mathsf{O}B$, *absent further information*.[61] But we may not want unrestricted detachment (or full modus ponens) for the conditional $\Rightarrow$. For instance, given the premises $p, q, p \Rightarrow \mathsf{O}r$, and $q \Rightarrow \mathsf{O}\neg r$, we may not want to be able to detach both $\mathsf{O}r$ and $\mathsf{O}\neg r$, unless perhaps we move to a non-standard characterization of O. So if we stick to a standard characterization of O as an **SDL**-operator, we will want to allow for some, but not all instances of modus ponens for $\Rightarrow$.

In other words, we only want to apply detachment in a defeasible

[60] One may also represent the conditional obligation "If A, then it is obligatory that B" by $\mathsf{O}(A \Rightarrow B)$ or $\mathsf{O}A \Rightarrow \mathsf{O}B$. We will have little to say about the first of these two alternatives; we briefly return to the second at the end of this section.

[61] We consider deontic detachment at the end of this section.

way. This can be done as follows in terms of ALs. We first enrich the language of **SDL** with a default conditional, where nested occurrences of $\Rightarrow$ are disallowed:

$$\begin{aligned}\mathcal{W}^{\Rightarrow} := \quad & \mathcal{W}^d \mid \langle \mathcal{W}^d \rangle \Rightarrow \langle \mathcal{W}^d \rangle \mid \neg\langle \mathcal{W}^{\Rightarrow} \rangle \mid \langle \mathcal{W}^{\Rightarrow} \rangle \vee \langle \mathcal{W}^{\Rightarrow} \rangle \mid \langle \mathcal{W}^{\Rightarrow} \rangle \wedge \\ & \langle \mathcal{W}^{\Rightarrow} \rangle \mid \\ & \langle \mathcal{W}^{\Rightarrow} \rangle \supset \langle \mathcal{W}^{\Rightarrow} \rangle \mid \langle \mathcal{W}^{\Rightarrow} \rangle \equiv \langle \mathcal{W}^{\Rightarrow} \rangle\end{aligned}$$

Next, let $\mathbf{SDL}_{\Rightarrow}$ be just **SDL**, but defined over this richer language. Hence, $\Rightarrow$ has no properties in $\mathbf{SDL}_{\Rightarrow}$. We then define our ALs on the basis of $\mathbf{SDL}_{\Rightarrow}$, by the set of abnormalities

$$\Omega_{\Rightarrow} =_{\mathsf{df}} \{(A \Rightarrow B) \wedge A \wedge \neg B \mid A, B \in \mathcal{W}^d\}$$

So whenever the conditional $A \Rightarrow B$ is true and A is true, then we assume that also B is true. Note that A and B can be arbitrary members of $\mathcal{W}^d$, hence also A can be a deontic statement such as $\mathsf{O}p$ – we return to this point below.

Let us call the resulting adaptive logics $\mathbf{SDL}^{\mathbf{x}}_{\Rightarrow}$. As the following proof illustrates, conditional obligations are detachable in $\mathbf{SDL}^{\mathbf{x}}_{\Rightarrow}$ as long as no conflicts are generated. (For the sake of readability, we abbreviate $(A \Rightarrow B) \wedge A \wedge \neg B$ as $A \not\Rightarrow B$).)

1	$p \wedge q$	Prem	$\emptyset$
2	$p \Rightarrow \mathsf{O}r$	Prem	$\emptyset$
3	$q \Rightarrow \mathsf{O}\neg r$	Prem	$\emptyset$
4	$(p \wedge q) \Rightarrow \mathsf{O}s$	Prem	$\emptyset$
5	$\mathsf{O}r$	1,2;RC	$\{p \not\Rightarrow \mathsf{O}r\}\checkmark^8$
6	$\mathsf{O}\neg r$	1,3;RC	$\{q \not\Rightarrow \mathsf{O}\neg r\}\checkmark^8$
7	$\mathsf{O}s$	1,4;RC	$\{(p \wedge q) \not\Rightarrow \mathsf{O}s\}$
8	$(p \not\Rightarrow \mathsf{O}r) \vee (q \not\Rightarrow \mathsf{O}\neg r)$	1-3;RU	$\emptyset$

The conditional $\Rightarrow$ of $\mathbf{SDL}_{\Rightarrow}$ is of course very weak – we can only make use of it by going adaptive. We can however strengthen the lower limit logic by adding further rules. Here are some candidates:

$$\text{If } A \Rightarrow C \text{ and } B \Rightarrow C, \text{ then } (A \vee B) \Rightarrow C \qquad \text{(Or)}$$

$$\text{If } A \Rightarrow B \text{ and } B \Rightarrow C, \text{ then } A \Rightarrow C \qquad \text{(Tra)}$$

$$\text{If } A \Rightarrow B \text{ and } (A \wedge B) \Rightarrow C, \text{ then } A \Rightarrow C \qquad \text{(CTra)}$$

$$\text{If } A \vdash B \text{ and } B \Rightarrow C, \text{ then } A \Rightarrow C \qquad \text{(SA)}$$

Each of these rules can be added to our logic if desired. However, one should be careful here, as adding more properties to one's lower limit logic often generates flip-flop problems, as explained in the previous sections of this chapter.

Unlike the dyadic deontic operator of $\mathbf{SDL_d}$ from Section 9.1, the conditional $\Rightarrow$ of $\mathbf{SDL}^{\mathbf{x}}_{\Rightarrow}$ is completely independent of the way we formalize obligations. We can read a statement $A \Rightarrow B$ as 'If A, then normally B' as we would do for defeasible conditionals in general. In $\mathbf{SDL}^{\mathbf{x}}_{\Rightarrow}$ we detach obligations via defeasible modus ponens, just like we defeasibly detach conclusions in default logic or in your preferred calculus of non-monotonic logic. So this approach is very unifying, treating deontic reasoning as just one specific type of defeasible reasoning in general.

However, the approach has the disadvantage that it cannot as easily accommodate deontic detachment (DD) (cf. Section 9.1). Consider the following three inferences:

$$p, p \Rightarrow \mathsf{O}q \vdash \mathsf{O}q \tag{83}$$

$$\mathsf{O}p, \mathsf{O}p \Rightarrow \mathsf{O}q \vdash \mathsf{O}q \tag{84}$$

$$\mathsf{O}p, p \Rightarrow \mathsf{O}q \vdash \mathsf{O}q \tag{85}$$

(83) and (84) are derivable $\mathbf{SDL}^{\mathbf{x}}_{\Rightarrow}$-rules: we can apply these rules conditionally in $\mathbf{SDL}^{\mathbf{x}}_{\Rightarrow}$. However, (85) is not a derivable rule in $\mathbf{SDL}^{\mathbf{x}}_{\Rightarrow}$. Some have argued that this is how it should be (see e.g. the discussion and references in [Bonevac, 2016]). Still, (85) has some intuitive force.

One way to defend $\mathbf{SDL}^{\mathbf{x}}_{\Rightarrow}$ is by arguing that, whenever we think deontic detachment should be allowed, the appropriate translation of the conditional is as in (84). More generally, such conditionals are of the form: if A is obligatory, then also B is obligatory ($\mathsf{O}A \Rightarrow \mathsf{O}B$). However, that would mean that in many cases we need a kind of "double translation" of deontic conditionals – as $(A \Rightarrow \mathsf{O}B) \wedge (\mathsf{O}A \Rightarrow \mathsf{O}B)$ – which seems highly artificial. Moreover, it would go against the spirit of the adaptive logic approach, where the idea is that the logic should determine which applications of deontic detachment are rational. So altogether, it seems that the second approach is less suited to accommodate (DD).

Further reading The literature on the formalization of defeasible conditionals is vast. For some good entry points, see e.g. [Kraus *et al.*, 1990; Makinson, 2005]. In this section we only presented a basic mechanism for the defeasible detachment of obligations via a new conditional. For more information on the types of rules that can be studied via this mechanism, we refer to [Straßer, 2014, Chapter 6].

9.3 Adaptive Characterizations of input/output logic

Input/output logic The third approach to deontic conditionals that we will discuss here goes under the name input/output logic (henceforth

I/O logic). Technically speaking, I/O logics (without constraints, cf. infra) are operations that map every pair $\langle \mathcal{A}, \mathcal{G} \rangle$ to an "output" $\mathcal{O} \subseteq \mathcal{W}$, where (i) $\mathcal{G} \subseteq \mathcal{W} \times \mathcal{W}$ is a set of "input/output pairs" (A, B); (ii) $\mathcal{A} \subseteq \mathcal{W}$ is the "input". For instance, given the input $\mathcal{A} = \{p, q\}$ and the set of conditionals $\mathcal{G} = \{(p, r), (q, s)\}$, the output $\mathcal{O}$ will consist of r, s, and everything that follows from their conjunction.

In a deontic setting, $\mathcal{A}$ usually represents factual information, $\mathcal{G}$ is a set of conditional obligations, and the output consists of what is obligatory, given the facts at hand and given the conditional obligations that make up our normative system. The idea of factual detachment thus lies at the very core of I/O-logics.

Different I/O-logics are obtained by varying on the rules under which $\mathcal{G}$ is closed, before one applies factual detachment. These rules are themselves highly similar to the ones used to characterize default conditionals (cf. Section 9.2). For example, by assuming that $\mathcal{G}$ is closed under the rule (OR)

$$\text{If } (A, C) \text{ and } (B, C), \text{ then } (A \vee B, C) \tag{OR}$$

we can obtain r in the output of $\mathcal{A} = \{p \vee q\}$ and $\mathcal{G} = \{(p, r), (q, r)\}$. Similarly, if $\mathcal{G}$ is closed under the rule (Tra), one can validate deontic detachment (DD):

$$\text{If } (A, B) \text{ and } (B, C), \text{ then } (A, C) \tag{Tra}$$

So for instance, given closure under (Tra), we can obtain q in the output of $\mathcal{A} = \emptyset$ and $\mathcal{G} = \{(\top, p), (p, q)\}$.

Both (FD) and (DD) are accommodated within the I/O-systems presented [Makinson and van der Torre, 2000]. However, this framework cannot handle conflicts that arise from the application of (FD) or (DD) or both: e.g. $\mathcal{A} = \{p, q\}$ and $\mathcal{G} = \{(p, r), (q, \neg r)\}$ will generate a trivial output.

To deal with such cases, Makinson and van der Torre introduced a set $\mathcal{C}$ of "constraints" in their [2001]. Depending on the application context $\mathcal{C}$ may represent physical constraints, human rights, practical considerations, etc. $\mathcal{C}$ can restrict the output in two ways, each corresponding to a different style of reasoning. We can require consistency of $\mathcal{O} \cup \mathcal{C}$, or we can impose the weaker requirement that for each $A \in \mathcal{O}$, $\{A\} \cup \mathcal{C}$ is consistent. In the border case where $\mathcal{C} = \emptyset$, this simply means that we require the $\mathcal{O}$ to be consistent, or that each $A \in \mathcal{O}$ is consistent. The first approach is called *meet* constrained output; the second is the *join* constrained output.

The adaptive characterization In [Straßer *et al.*, 2016], I/O-logics are characterized in terms of deductive systems within a rich modal language. We explain how this works for constrained I/O-logics (the case for unconstrained I/O-logics is simpler). The language uses unary modal operators $\mathsf{in}, \mathsf{out}, \mathsf{con}$ to represent input, output, and constraints respectively. Input/output pairs (A, B) are represented by means of $\mathsf{in}, \mathsf{out}$ and a conditional $\rightarrow$, as follows:

$$\mathsf{in}A \rightarrow \mathsf{out}B$$

The principle of detachment and the rules for input/output-pairs are then translated into the object level. This gives us rules and axioms such as the following:

$$\text{If } \mathsf{in}A \text{ and } \mathsf{in}A \rightarrow \mathsf{out}B, \text{ then } \mathsf{out}B \tag{DET$'$}$$
$$((\mathsf{in}A \rightarrow \mathsf{out}C) \wedge (\mathsf{in}B \rightarrow \mathsf{out}C)) \supset (\mathsf{in}(A \vee B) \rightarrow \mathsf{out}C) \tag{OR$'$}$$
$$((\mathsf{in}A \rightarrow \mathsf{out}B) \wedge (\mathsf{in}B \rightarrow \mathsf{out}C)) \supset (\mathsf{in}A \rightarrow \mathsf{out}C) \tag{Tra$'$}$$

The fact that the output should be consistent with the set of constraints is captured by

$$\mathsf{con}A \supset \neg\mathsf{out}\neg A \tag{ROC}$$

Finally, to mimic the selection of maximal consistent sets of conditionals, a dummy operator $\bullet$ is introduced and used in much the same way as we did in Section 3. That is, conditionals $(A, B) \in \mathcal{G}$ are translated into formulas of the form $\bullet(\mathsf{in}A \rightarrow \mathsf{out}B)$. The adaptive logics then allow one to "activate" such conditionals by removing the dummy, whence one can apply rules like (DET$'$), (OR$'$), or (Tra$'$) to them.

Suppose, for instance, that we are given the following set of inputs, I/O-pairs, and constraints: $\mathcal{A} = \{p, q\}, \mathcal{G} = \{(p, r), (q, s), (p, t)\}, \mathcal{C} = \{\neg r \vee \neg s\}$. In the language from [Straßer *et al.*, 2016], this gives us the following premise set:

$$\Gamma = \{\mathsf{in}p, \mathsf{in}q, \bullet(\mathsf{in}p \rightarrow \mathsf{out}r), \bullet(\mathsf{in}q \rightarrow \mathsf{out}s), \bullet(\mathsf{in}p \rightarrow \mathsf{out}t), \mathsf{con}(\neg r \vee \neg s)\}$$

In an adaptive proof from Γ, we can finally derive $\mathsf{out}t$. Depending on the strategy, we can also finally derive $\mathsf{out}(r \vee s)$ or even $\mathsf{out}r$ and $\mathsf{out}s$.

Let us illustrate this with an object-level proof. To enhance readibility, we use $\star(A, B)$ to abbreviate $\bullet(\mathsf{in}A \rightarrow \mathsf{out}B) \wedge \neg(\mathsf{in}A \rightarrow \mathsf{out}B)$.

Moreover, we use superscripts r, m to indicate the strategy under which certain lines are (not) marked:[62]

1	$\mathsf{in}p$	Prem	$\emptyset$
2	$\mathsf{in}q$	Prem	$\emptyset$
3	$\bullet(\mathsf{in}p \rightarrow \mathsf{out}r)$	Prem	$\emptyset$
4	$\bullet(\mathsf{in}q \rightarrow \mathsf{out}s)$	Prem	$\emptyset$
5	$\bullet(\mathsf{in}p \rightarrow \mathsf{out}t)$	Prem	$\emptyset$
6	$\mathsf{con}(\neg r \vee \neg s)$	Prem	$\emptyset$
7	$\mathsf{in}p \rightarrow \mathsf{out}r$	3; RC	$\{\star(p,r)\}\checkmark^{\mathsf{r},\mathsf{m}}$
8	$\mathsf{in}q \rightarrow \mathsf{out}s$	4; RC	$\{\star(q,s)\}\checkmark^{\mathsf{r},\mathsf{m}}$
9	$\mathsf{in}p \rightarrow \mathsf{out}t$	5; RC	$\{\star(p,t)\}$
10	$\mathsf{out}r$	1,7; RU	$\{\star(p,r)\}\checkmark^{\mathsf{r},\mathsf{m}}$
11	$\mathsf{out}s$	2,8; RU)	$\{\star(q,s)\}\checkmark^{\mathsf{r},\mathsf{m}}$
12	$\mathsf{out}t$	1,9; RU	$\{\star(p,t)\}$
13	$\mathsf{out}r \vee \mathsf{out}s$	10; RU	$\{\star(p,r)\}\checkmark^{\mathsf{r}}$
14	$\mathsf{out}r \vee \mathsf{out}s$	11; RU	$\{\star(q,s)\}\checkmark^{\mathsf{r}}$
15	$\star(p,r) \vee \star(q,s)$	1-4,6; RU	$\emptyset$

Under the modal translation, the minimal abnormality strategy corresponds to the operation of meet constrained output; normal selections (cf. Section 3.4 corresponds to the join constrained output. The reliability strategy has no counterpart in the original framework of [Makinson and van der Torre, 2001]; however, as shown in [Straßer *et al.*, 2016], one can also define a procedural semantics for the corresponding operation, much in the spirit of Makinson and van der Torre's original setting.

Further reading I/O-logic was introduced by Makinson and van der Torre [2000; 2001] as a formal tool for modeling non-monotonic reasoning with conditionals. We refer to [Parent and van der Torre, 2013] in the first volume of this handbook for an introduction to this approach and its applications to deontic reasoning.

The framework presented here is not only sufficient to characterize many well-known I/O logics, but it allows one to go beyond the expressive means of I/O logics so as to express useful notions in deontic logic such as violations and sanctions. We refer to [Straßer *et al.*, 2016] for the many details, and for an elaborate presentation and discussion of these advantages.

[62] The formulas at lines 10-12 are derivable in view of (DET′). The formula at line 15 is derivable in view of (DET′), modal properties of the **KD**-operator out, and the axiom schema (ROC).

10 Deontic compatibility

10.1 Adaptive logics for deontic compatibility

We saw how ALs are useful for reasoning in the presence of normative conflicts, and for detaching conditional obligations. A different context of application for ALs that was mentioned in Section 1 concerns the implementation of the *nullum crimen sine lege* principle (henceforth NCSL). This principle expresses that no crimes occur where there is no law: that which is not forbidden, is permitted. Typically, NCSL is understood as a rule of closure permitting all the actions not prohibited by penal law [Alchourrón and Bulygin, 1971, pp. 142–143]. It is a fundamental principle of law, the roots of which go back at least as far as the French Revolution. In the twentieth century it was incorporated in various human rights instruments as a non-derogable right [Mokhtar, 2005].

Logicians and computer scientists are very familiar with the concept of "negation by default", according to which a piece of information represented by some variable is taken to be absent unless and until we include it in our database. For instance, where a variable x abbreviates that there is a train leaving for Ghent at 14:14, we may conclude that $\neg x$ unless x is mentioned on the timetable at the train station. Similarly, we can think of NCSL as "permission by default". Formally, this can be expressed as follows, where we take our premise set Γ to represent a given normative system or law, and where $\vdash$ is an ordinary (Tarskian) deontic logic:

$$\Gamma \vdash \mathsf{P}A \text{ iff } \Gamma \nvdash \neg\mathsf{P}A$$

Assume that we want to implement this equivalence against the background of full **SDL**. Then, on pain of inconsistency, the equivalence can at best hold *defeasibly*. Suppose, for instance, that we are given a premise set Γ such that $\Gamma \vdash \neg\mathsf{P}p \vee \neg\mathsf{P}q$, while $\Gamma \nvdash \neg\mathsf{P}p$ and $\Gamma \nvdash \neg\mathsf{P}q$. Then we cannot preserve consistency *and* apply NCSL to derive $\mathsf{P}p$ as well as $\mathsf{P}q$. What we want, then, is a logic that preserves consistency and applies NCSL *as much as possible*.

This motivates an adaptive logic of deontic compatibility which implements NCSL by taking **SDL** as its lower limit logic, and Ω_{P} as its set of abnormalities:

$$\Omega_{\mathsf{P}} = \{\neg\mathsf{P}A \mid A \in \mathcal{W}\}$$

We call the resulting logic $\mathbf{SDL}^{\mathsf{x}}_{\mathbf{nc}}$ with nc for *nullum crimen* and $\mathsf{x} \in \{\mathsf{r},\mathsf{m}\}$. In view of the **SDL**-validity of $\mathsf{P}A \vee \neg\mathsf{P}A$, $\mathbf{SDL}^{\mathsf{x}}_{\mathbf{nc}}$ allows for the inference of jointly compatible permissions relative to a given premise set. The following object level proof further illustrates the ways this logic works.

1	$\mathsf{O}(\neg p \vee \neg q)$	Prem	$\emptyset$
2	$\mathsf{O}(\neg s \wedge t)$	Prem	$\emptyset$
3	$\mathsf{P}t \supset (\mathsf{P}u \supset \mathsf{O}\neg v)$	Prem	$\emptyset$
4	$\mathsf{P}p$	RC	$\{\neg\mathsf{P}p\}$
5	$\mathsf{P}\neg p$	RC	$\{\neg\mathsf{P}\neg p\}$
6	$\mathsf{P}q$	RC	$\{\neg\mathsf{P}q\}$
7	$\mathsf{P}\neg q$	RC	$\{\neg\mathsf{P}\neg q\}$
8	$\mathsf{P}r$	RC	$\{\neg\mathsf{P}r\}$
9	$\mathsf{P}\neg r$	RC	$\{\neg\mathsf{P}\neg r\}$
10	$\mathsf{P}s$	RC	$\{\neg\mathsf{P}s\}\checkmark^{18}$
11	$\mathsf{P}\neg s$	2; RU	$\emptyset$
12	$\mathsf{P}t$	2; RU	$\emptyset$
13	$\mathsf{P}\neg t$	RC	$\{\neg\mathsf{P}\neg t\}\checkmark^{19}$
14	$\mathsf{P}u$	RC	$\{\neg\mathsf{P}u\}\checkmark^{20}$
15	$\mathsf{P}\neg u$	RC	$\{\neg\mathsf{P}\neg u\}$
16	$\mathsf{P}v$	RC	$\{\neg\mathsf{P}v\}\checkmark^{20}$
17	$\mathsf{P}\neg v$	RC	$\{\neg\mathsf{P}\neg v\}$
18	$\neg\mathsf{P}s$	2; RU	$\emptyset$
19	$\neg\mathsf{P}\neg t$	2;RU	$\emptyset$
20	$\neg\mathsf{P}u \vee \neg\mathsf{P}v$	2;3;RU	$\emptyset$

One nice feature of this logic is its simplicity, when restricted to premise sets of the form $\{\mathsf{O}A \mid A \in \Delta\}$ for $\Delta \subseteq \mathcal{W}$. Indeed, for such cases, the strategies *reliability* and *minimal abnormality* will coincide, since every minimal Dab-consequence of such premise sets contains only one disjunct $A \in \Omega_{\mathsf{P}}$. This is itself an immediate corollary of the following:

Proposition 10.1. *If* $\Gamma = \{\mathsf{O}A \mid A \in \Delta\}$ *for* $\Delta \subseteq \mathcal{W}$*, then* $\Gamma \vdash_{\mathbf{SDL}} (\neg\mathsf{P}A_1 \vee \ldots \vee \neg\mathsf{P}A_n)$ *iff there is an* $i \in \{1,\ldots,n\}$ *such that* $\Gamma \vdash_{\mathbf{SDL}} \neg\mathsf{P}A_i$.

In more complex cases such as our example proof above, the two strategies may well differ. In either case, the resulting consequence set will be closed under **SDL** and consistent.

One may wonder whether the idea of deontic compatibility should necessarily be phrased in terms of the underlying logic **SDL** – after all, legal conflicts are a fact of life, and as soon as such conflicts are modeled

in **SDL**, everything becomes obligatory and permissible. This motivates a logic that defeasibly applies NCSL *and* that accommodates conflicts much as the logics presented in Sections 5-7.

Let us illustrate this by means of the paraconsistent deontic logics from Section 7. One option is to just take a monotonic paraconsistent deontic logic – say **DCLuN**, to keep things relatively simple – and to use as a set of abnormalities

$$\Omega = \{\mathsf{O}A \mid A \in \mathcal{W}^{\sim}\}$$

However, the resulting logic will be too strong, in the sense that it will allow one to derive permissions that should intuitively not be derivable, even if we take NCSL seriously. With such a logic, one can e.g. derive $\mathsf{P}^{\neg}_{\neg}{\sim}p$ from $\Gamma = \{\mathsf{O}p\}$. The underlying reason is that in these logics, $\mathsf{O}p$ does not entail $\mathsf{O}\neg{\sim}p$ (just like the truth of p does not entail the falsehood of ${\sim}p$ in their paraconsistent propositional base), and hence one can consistently assume that $\mathsf{O}\neg{\sim}p$ is false even when $\mathsf{O}p$ is true. But the mere fact that we want to allow for the logical possibility of conflicts, should not entail that everything is permissible.

A more plausible combination of conflict-tolerance and *nullum crimen* can be obtained if we combine the *adaptive* logics $\mathbf{DCLuN^x}$ from Section 7 with NCSL, using the format of lexicographic ALs that was introduced in Section 3.4. This means that the logic first minimizes inconsistencies (which implies i.a. that we derive further obligations), and only after that do we maximize permissions. In this way we can e.g. explain why in view of $\Gamma' = \{\mathsf{O}p, \mathsf{O}({\sim}p \vee q), \mathsf{O}r, \mathsf{O}{\sim}r\}$ we can derive $\mathsf{O}q$, $\mathsf{O}p$ and $\neg\mathsf{P}^{\neg}_{\neg}{\sim}p$, $\neg\mathsf{P}^{\neg}_{\neg}{\sim}q$, but also $\mathsf{P}^{\neg}_{\neg}s, \mathsf{P}^{\neg}_{\neg}{\sim}s$, and $\mathsf{P}^{\neg}_{\neg}r, \mathsf{P}^{\neg}_{\neg}{\sim}r$.

Analogously, one may enrich the logics from Sections 5 and 6 with a default version of NCSL. For similar reasons as in the paraconsistent case, it seems best to first apply the adaptive mechanisms from those sections, and only after that to apply NCSL. For instance, in the case of non-aggregative deontic logics, we would not want to infer $\mathsf{P}\neg(p{\wedge}q)$ from $\Gamma = \{\mathsf{O}p, \mathsf{O}q\}$. Likewise, in the context of the **LUM**-logics, we would not want to infer $\mathsf{P}\neg p$ from $\Gamma' = \{\mathsf{O}(p \wedge q)\}$. The full development of such rich ALs for deontic compatibility is still very much open; it should by now be clear that a broad range of options are to be considered, and that the devil may well be in the many details.

10.2 Further reading

Adaptive logics for classical compatibility were among the first ampliative adaptive logics to be published – see [Batens and Meheus, 2000].

Although these logics were not formulated in the standard format, one can do this by means of the triple

$$\langle \mathbf{S5}, \{\neg\Diamond A \mid A \text{ is a non-modal formula } \}, \mathsf{x} \in \{\mathsf{r}, \mathsf{m}\}\rangle$$

The relation between classical compatibility and the logics in question is then expressed in terms of a modal translation: A is compatible with Γ iff $\{\Box A \mid A \in \Gamma\} \vdash_{\mathbf{AL}} \Diamond A$.

In [Meheus, 2003], the basic idea behind these logics is used in order to develop a formal account of paraconsistent compatibility, i.e., what it means that a given formula is compatible with a certain (possibly inconsistent) scientific theory. As Meheus argues there, one also first needs to minimize inconsistencies before checking compatibility with the resulting maximally consistent interpretation of the theory.

11 Summary and outlook

This chapter started with two simple adaptive logics that can handle deontic conflicts. We then discussed in some detail more sophisticated conflict-tolerant ALs, as well as ALs for reasoning with conditional obligations and the problems of detachment that are associated with these. Finally, we broadened the picture by presenting ALs for the inherently defeasible *nullum crimen sine lege* principle. This should convince the reader of the generality and the flexibility of the adaptive logic framework.

It is important to realize, however, that this does not exhaust the possibilities of adaptive logics for the domain of normative reasoning. This requires more explanation.

All logics presented in this chapter share important constraints. One of them is that we only considered the two main deontic modalities, "it is obligatory that" and "it is permitted that", and we moreover restricted our formal languages to non-nested occurrences of those modalities. Another one is that we took it for granted that we can start from premise sets that merely consist of very specific and very concrete normative statements, like "Nathan ought to take Lisa to that particular movie on Saturday afternoon".

Because of these constraints, the logics allow us to explicate only a very small part of the normative reasoning one finds in actual cases. Already the everyday examples from Nathan's life (that are recognizable to many of us) suffice to illustrate this. In Nathan's first predicament (the preludium), his normative reasoning does not start from the statements

that he ought to take Lisa to the movie in the afternoon, that he ought to look after Ben in the afternoon and that he ought to take Lisa for a veggie burger in the evening. These statements are themselves *derived* from other statements, in this case concrete promises by Nathan and the general rule "One ought to keep one's promises". Also in Nathan's second predicament (Section 3.1), the specific normative statements are not given at the outset, but are the result of reasoning. In this case, not only general rules play a role (like "One ought to return favors"), but also commands uttered by an authority (i.c. Nathan's father). None of the logics presented here allows us to explicate the reasoning from general rules to their instances or from commands (uttered by one person) to obligations (for another person) – to mention only two possible origins of specific normative statements.

There is more. Some readers may have noticed that, while presenting our conflict-tolerant logics, we used the term "*prima facie* obligations", but never used the term "all-things-considered obligations" which is, at least since Ross' [1930], associated with it. Instead we consistently used the term "actual obligations". The reason is that none of our logics enables us to explicate the reasoning from *prima facie* obligations to all-things-considered obligations, where the latter is taken to mean something like "obligations that are, after careful deliberation, considered to be binding". Our logics only give us those binding obligations for which relatively little deliberation is needed. For instance, "if a *prima facie* obligation is unconflicted, it should be binding" or "if two *prima facie* obligations are unconflicted, also their conjunction should be binding", etc.

In order to explicate the reasoning that goes on in resolving a predicament and finding out what one's all-things-considered obligations are (or should be), we need much more than just deontic operators. For instance, whatever Nathan's solution for his first predicament may be, it will involve certain beliefs (for instance, what Nathan believes will happen if he does not keep the promise he made to his mother). None of our logics can handle interactions between deontic modalities on the one hand, and doxastic or epistemic modalities on the other.[63]

Does this mean we have gone all this way for nothing? Certainly not. We are convinced that the logics presented here are good candidates to explicate *part of* the reasoning that goes on in specific deontic contexts. They moreover provide a first stepping stone to more complex, richer

[63] See e.g. [Pacuit *et al.*, 2006] for a study of the interaction between epistemic and deontic modalities.

accounts of deontic reasoning. So there is still hope for Nathan, or at least for us to fully understand how he should reason.

References

[Alchourrón and Bulygin, 1971] Carlos E. Alchourrón and Eugenio Bulygin. *Normative Systems*. Springer-Verlag, Wien/New York, 1971.

[Allo, 2013] Patrick Allo. Adaptive logic as a modal logic. *Studia Logica*, 101(5):933–958, 2013.

[Åqvist, 2002] Lennart Åqvist. Deontic logic. In Dov Gabbay and Franz Guenthner, editors, *Handbook of Philosophical Logic (2nd edition)*, volume 8, pages 147–264. Kluwer Academic Publishers, 2002.

[Batens and Meheus, 2000] Diderik Batens and Joke Meheus. The adaptive logic of compatibility. *Studia Logica*, 66:327–348, 2000.

[Batens *et al.*, 2009] Diderik Batens, Christian Straßer, and Peter Verdée. On the transparency of defeasible logics: Equivalent premise sets, equivalence of their extensions, and maximality of the lower limit. *Logique et Analyse*, 207:281–304, 2009.

[Batens, 1986] D. Batens. Dialectical dynamics within formal logics. *Logique et Analyse*, 114:161–173, 1986.

[Batens, 1989] D. Batens. Dynamic dialectical logics. In G. Priest, R. Routley, and J. Norman, editors, *Paraconsistent Logic. Essays on the Inconsistent*, pages 187–217. Philosophia Verlag, München, 1989.

[Batens, 1997] Diderik Batens. Inconsistencies and beyond. A logical-philosophical discussion. *Revue Internationale de Philosophie*, 200:259–273, 1997.

[Batens, 1999a] Diderik Batens. Inconsistency-adaptive logics. In Ewa Orłowska, editor, *Logic at Work. Essays dedicated to the memory of Helena Rasiowa*, pages 445–472. Physica Verlag (Springer), Heidelberg, New York, 1999.

[Batens, 1999b] Diderik Batens. Zero logic adding up to classical logic. *Logical Studies*, 2:15, 1999.

[Batens, 2001] Diderik Batens. A general characterization of adaptive logics. *Logique et Analyse*, 173–175:45–68, 2001. Appeared 2003.

[Batens, 2007] Diderik Batens. A universal logic approach to adaptive logics. *Logica Universalis*, 1:221–242, 2007.

[Batens, 2011] Diderik Batens. Logics for qualitative inductive generalization. *Studia Logica*, 97:61–80, 2011.

[Batens, 2015] Diderik Batens. Tutorial on inconsistency-adaptive logics. In Jean-Yves Béziau, Mihir Chakraborty, and Soma Dutta, editors, *Springer Proceedings in Mathematics & Statistics*, volume 152, pages 3–38. Springer, 2015.

[Beirlaen and Aliseda, 2014] Mathieu Beirlaen and Atocha Aliseda. A conditional logic for abduction. *Synthese*, 191(15):3733–3758, 2014.

[Beirlaen and Straßer, 2011] Mathieu Beirlaen and Christian Straßer. A paraconsistent multi-agent framework for dealing with normative conflicts. In Joao Leite, Paolo Torroni, Thomas Agotnes, Guido Boella, and Leon van der Torre, editors, *Computational Logic in Multi-Agent Systems*, volume 6814 of *Lecture Notes in Computer Science*, pages 312–329. Springer, Berlin/Heidelberg, 2011.

[Beirlaen and Straßer, 2013] M. Beirlaen and C. Straßer. Two adaptive logics of norm-propositions. *Journal of Applied Logic*, 11(2):147–168, 2013.

[Beirlaen and Straßer, 2014] M. Beirlaen and C. Straßer. Nonmonotonic reasoning with normative conflicts in multi-agent deontic logic. *Journal of Logic and Computation*, 24:1179–1207, 2014.

[Beirlaen and Straßer, 2016] M. Beirlaen and C. Straßer. A structured argumentation framework for detaching conditional obligations. In O. Roy, A. Tamminga, and M. Willer, editors, *Proceedings of the 13th International Conference on Deontic Logic and Normative Systems (ΔEON 2016, Bayreuth, Germany)*, pages 32–48. College Publications, 2016.

[Beirlaen *et al.*, 2013] Mathieu Beirlaen, Christian Straßer, and Joke Meheus. An inconsistency-adaptive deontic logic for normative conflicts. *Journal of Philosophical Logic*, 42(2):285–315, 2013.

[Beirlaen *et al.*, 2018] Mathieu Beirlaen, Bert Leuridan, and Frederik Van De Putte. A logic for the discovery of deterministic causal regularities. *Synthese*, 195:367–399, 2018.

[Beirlaen, 2012] M. Beirlaen. *Tolerating Normative Conflicts in Deontic Logic*. Dissertation, Ghent University, 2012. Available online at `http://www.clps.ugent.be/research/doctoral-dissertations`.

[Belnap and Perloff, 1993] N. Belnap and M. Perloff. In the realm of agents. *Annals of Mathematics and Artificial Intelligence*, 9:25–48, 1993.

[Blackburn *et al.*, 2001] Patrick Blackburn, Maarten De Rijke, and Yde Venema. *Modal Logic*. Cambridge Tracts in Theoretical Computer Science, 2001.

[Bonevac, 2016] Daniel Bonevac. Defaulting on reasons. *Noûs*, 2016.

[Brink, 1994] D. Brink. Moral conflict and its structure. *The Philosophical Review*, 103:215–247, 1994.

[Cariani, 2013] Fabrizio Cariani. "Ought" and resolution semantics. *Noûs*, 47(3):534–558, 2013.

[Carmo and Jones, 2002] J. Carmo and A. Jones. Deontic logic and contrary-to-duties. In Dov Gabbay and Franz Guenthner, editors, *Handbook of Philosophical Logic (2nd edition)*, volume 8, pages 265–343. Kluwer Academic Publishers, 2002.

[Chellas, 1980] Brian Chellas. *Modal Logic: an Introduction*. Cambridge: Cambridge university press, 1980.

[Chisholm, 1963] R. Chisholm. Contrary-to-duty imperatives and deontic

logic. *Analysis*, 27:33–36, 1963.

[Elgesem, 1997] Dag Elgesem. The modal logic of agency. *Nordic Journal of Philosophical Logic*, 2(2):1–46, 1997.

[Fine, 2016] Kit Fine. Angellic content. *Journal of Philosophical Logic*, 45(2):199–226, April 2016.

[Gabbay, 2012] Dov M. Gabbay. Bipolar argumentation frames and contrary to duty obligations, preliminary report. In M. Fisher, L. van der Torre, M. Dastani, and G. Governatori, editors, *Computational Logic in Multi-Agent Systems*, pages 1–24. Springer, 2012.

[Goble, 1990a] L. Goble. A logic of "good", "should", and "would": Part I. *Journal of Philosophical Logic*, 19:169–199, 1990.

[Goble, 1990b] Lou Goble. A logic of "good", "should", and "would": Part II. *Journal of Philosophical Logic*, 19:253–76, 1990.

[Goble, 2000] Lou Goble. Multiplex semantics for deontic logic. *Nordic Journal of Philosophical Logic*, 5:113–134, 2000.

[Goble, 2003] Lou Goble. Preference semantics for deontic logic. Part I: Simple models. *Logique et Analyse*, 183–184:383–418, 2003.

[Goble, 2004a] L. Goble. A proposal for dealing with deontic dilemmas. In A. Lomuscio and D. Nute, editors, *7th International Workshop on Deontic Logic in Computer Science*, volume 3065 of *Lecture Notes in Computer Science*, pages 74–113. Springer, 2004.

[Goble, 2004b] Lou Goble. Preference semantics for deontic logic. Part II: Multiplex models. *Logique et Analyse*, 185–188:335–363, 2004.

[Goble, 2005] Lou Goble. A logic for deontic dilemmas. *Journal of Applied Logic*, 3:461–483, 2005.

[Goble, 2009] L. Goble. Normative conflicts and the logic of ought. *Noûs*, 43:450–489, 2009.

[Goble, 2013] Lou Goble. Prima facie norms, normative conflicts, and dilemmas. In Dov Gabbay, Leon van der Torre, John Horty, and Xavier Parent, editors, *Handbook of Deontic Logic and Normative Systems*, volume 1, chapter 4, pages 241–351. College Publications, 2013.

[Goble, 2014] Lou Goble. Deontic logic (adapted) for normative conflicts. *Logic Journal of the IGPL*, 22(2):206–235, 2014.

[Gowans, 1987] C.W. Gowans, editor. *Moral Dilemmas*. Oxford University Press, 1987.

[Hansen, 2013] J. Hansen. Imperative logic and its problems. In Dov Gabbay, Leon van der Torre, John Horty, and Xavier Parent, editors, *Handbook of Deontic Logic and Normative Systems*, volume 1, chapter 2, pages 137–192. College Publications, 2013.

[Hansen, 2014] Jörg Hansen. *Reasoning about permission and obligation*, volume 3 of *Outstanding contributions to logic*, chapter 14, pages 287–333. Springer, 2014.

[Hansson, 1969] Bengt Hansson. An analysis of some deontic logics. *Noûs*,

3:373–398, 1969.

[Heyninck and Straßer, 2016] Jesse Heyninck and Christian Straßer. Relations between assumption-based approaches in nonmonotonic logic and formal argumentation. In Gabriele Kern-Isberner and Renata Wassermann, editors, *16th International Workshop on Non-Monotonic Reasoning, Cape Town, South Africa*, pages 65–76, 2016.

[Hilpinen and McNamara, 2013] Risto Hilpinen and Paul McNamara. Deontic logic: a historical survey and introduction. In Dov Gabbay, Leon van der Torre, John Horty, and Xavier Parent, editors, *Handbook of Deontic Logic and Normative Systems*, volume 1, chapter 1, pages 3–136. College Publications, 2013.

[Horty, 1994] J. Horty. Moral dilemmas and nonmonotonic logic. *Journal of Philosophical Logic*, 23(1):35–66, 1994.

[Horty, 1997] J. Horty. Nonmonotonic foundations for deontic logic. In Donald Nute, editor, *Defeasible Deontic Logic: Essays in Nonmonotonic Normative Reasoning*, pages 17–44. Kluwer Academic Publishers, 1997.

[Horty, 2002] J. Horty. Skepticism and floating conclusions. *Artificial Intelligence*, 135:55–72, 2002.

[Horty, 2003] J. Horty. Reasoning with moral conflicts. *Noûs*, 37:557–605, 2003.

[Horty, 2012] J. Horty. *Reasons as Defaults*. Oxford University Press, 2012.

[Jackson, 1985] Frank Jackson. On the semantics and logic of obligation. *Mind*, 94:177–195, 1985.

[Kraus *et al.*, 1990] S. Kraus, D. Lehmann, and M. Magidor. Nonmonotonic reasoning, preferential models and cumulative logics. *Artificial Intelligence*, 44:167–207, 1990.

[Leuridan, 2009] Bert Leuridan. Causal discovery and the problem of ignorance. An adaptive logic approach. *Journal of Applied Logic*, 7(2):188–205, 2009.

[Makinson and Schlechta, 1991] D. Makinson and K. Schlechta. Floating conclusions and zombie paths: two deep difficulties in the "directly skeptical" approach to defeasible inheritance nets. *Artificial Intelligence*, 48:199–209, 1991.

[Makinson and van der Torre, 2000] D. Makinson and L. van der Torre. Input/output logics. *Journal of Philosophical Logic*, 29:383–408, 2000.

[Makinson and van der Torre, 2001] D. Makinson and L. van der Torre. Constraints for input/output logics. *Journal of Philosophical Logic*, 30:155–185, 2001.

[Makinson, 1994] David Makinson. General patterns in nonmonotonic reasoning. In *Handbook of Logic in Artificial Intelligence and Logic Programming, vol. III*. Clarendon Press, 1994.

[Makinson, 2005] David Makinson. *Bridges from Classical to Nonmonotonic Logic*, volume 5 of *Texts in Computing*. King's College Publications, London, 2005.

[Marcus, 1980] Ruth Barcan Marcus. Moral dilemmas and consistency. *Journal of Philosophy*, 77:121–136, 1980. Reprinted in [Gowans, 1987].

[McGinnis, 2007a] C. McGinnis. *Paraconsistency and Deontic Logic: Formal Systems for Reasoning with Normative Conflicts*. Dissertation, University of Minnesota, 2007.

[McGinnis, 2007b] C. McGinnis. Semi-paraconsistent deontic logic. In Jean-Yves Béziau, Walter Carnielli, and Dov Gabbay, editors, *Handbook of Paraconsistency*, pages 81–99. College Publications, London, 2007.

[Meheus *et al.*, 2002] Joke Meheus, Liza Verhoeven, Maarten Van Dyck, and Dagmar Provijn. Ampliative adaptive logics and the foundation of logic-based approaches to abduction. In L. Magnani, N.J. Nersessian, and Claudio Pizzi, editors, *Logical and Computational Aspects of Model-Based Reasoning*, pages 39–71. Kluwer Academic, Dordrecht, 2002.

[Meheus *et al.*, 2010] Joke Meheus, Mathieu Beirlaen, and Frederik Van De Putte. Avoiding deontic explosion by contextually restricting aggregation. In Guido Governatori and Giovanni Sartor, editors, *Deontic Logic in Computer Science*, volume 6181 of *Lecture Notes in Computer Science*, pages 148–165. Springer Berlin Heidelberg, 2010.

[Meheus *et al.*, 2012] Joke Meheus, Mathieu Beirlaen, Frederik Van De Putte, and Christian Straßer. Non-adjunctive deontic logics that validate aggregation as much as possible, 2012. Unpublished manuscript, available at `http://www.clps.ugent.be/research/publications`.

[Meheus *et al.*, 2016] Joke Meheus, Christian Straßer, and Peter Verdée. Which style of reasoning to choose in the face of conflicting information? *Journal of Logic and Computation*, 26(1):361–380, 2016.

[Meheus, 2003] Joke Meheus. Paraconsistent compatibility. *Logique et Analyse*, 183–184:251–287, 2003.

[Mokhtar, 2005] Ali Mokhtar. Nullum crimen, nulla poena sine lege: Aspects and prospects. *Statute Law Review*, 26(1):41–55, 2005.

[Nute, 1999] D. Nute. Norms, priorities, and defeasibility. In Paul McNamara and Henri Prakken, editors, *Norms, Logics and Information Systems. New Studies on Deontic Logic and Computer Science*, pages 201–218. IOS Press, 1999.

[Odintsov and Speranski, 2013] Sergei Odintsov and Stanislav Speranski. Computability issues for adaptive logics in expanded standard format. *Studia Logica*, 101(6):1237–1262, 2013.

[Pacuit *et al.*, 2006] Eric Pacuit, Rohit Parikh, and Eva Cogan. The logic of knowledge based obligation. *Synthese*, 149(2):311–341, 2006.

[Parent and van der Torre, 2013] Xavier Parent and Leendert van der Torre. Input/output logic. In Dov Gabbay, Jeff Horty, Xavier Parent, Ron van der Meyden, and Leendert van der Torre, editors, *Handbook of Deontic Logic and Normative Systems*, volume 1, chapter 8, pages 499–544. College Publications, 2013.

[Parent and van der Torre, 2014] X. Parent and L. van der Torre. "Sing and

dance!" Input/output logics without weakening. In F. Cariani, D. Grossi, J. Meheus, and X. Parent, editors, *DEON (12th International Conference on Deontic Logic in Computer Science)*, volume 8554 of *Lecture Notes in Artificial Intelligence*, pages 149–165. Springer, 2014.

[Prakken and Sartor, 2015] H. Prakken and G. Sartor. Law and logic: A review from an argumentation perspective. *Artificial Intelligence*, 227:214–245, 2015.

[Prakken, 2002] Henri Prakken. Intuitions and the modelling of defeasible reasoning: some case studies. In *Proceedings of the Ninth International Workshop on Nonmonotonic Reasoning*, pages 91–99, Toulouse, 2002.

[Priest *et al.*, 2015] G. Priest, K. Tanaka, and Z. Weber. Paraconsistent logic. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Spring 2015 edition, 2015. http://plato.stanford.edu/archives/spr2015/entries/logic-paraconsistent/.

[Priest, 1987] Graham Priest. *In Contradiction. A Study of the Transconsistent*. Nijhoff, Dordrecht, 1987.

[Priest, 2002] G. Priest. Paraconsistent logic. In D. Gabbay and F. Guenthner, editors, *Handbook of Philosophical Logic (2nd edition)*, volume 8, pages 287–393. Kluwer Academic Publishers, 2002.

[Rescher and Manor, 1970] N. Rescher and R. Manor. On inferences from inconsistent premises. *Theory and Decision*, 1:179–217, 1970.

[Ross, 1930] W. David Ross. *The Right and the Good*. Clarendon Press, 1930.

[Schotch and Jennings, 1981] Peter K. Schotch and Raymond E. Jennings. Non-kripkean deontic logic. In Risto Hilpinen, editor, *New Studies in Deontic Logic*, pages 149–162. Reidel, Dordrecht, 1981.

[Segerberg, 1971] K. Segerberg. An essay in classical modal logic, 1971.

[Segerberg, 1992] K. Segerberg. Getting started: beginnings in the logic of action. *Studia Logica*, 51:347–378, 1992.

[Straßer and Arieli, 2019] Christian Straßer and Ofer Arieli. Normative reasoning by sequent-based argumentation. *Journal of Logic and Computation*, 29(3):387–415, 2019.

[Straßer and Beirlaen, 2011] Christian Straßer and Mathieu Beirlaen. Towards more conflict-tolerant deontic logics by relaxing the interdefinability between obligations and permissions, 2011. Unpublished manuscript, available at `http://www.clps.ugent.be/research/publications`.

[Straßer and Šešelja, 2010] Christian Straßer and Dunja Šešelja. Towards the Proof-theoretic Unification of Dung's Argumentation Framework: an Adaptive Logic Approach. *Journal of Logic and Computation*, 21:133–156, 2010.

[Straßer *et al.*, 2012] Christian Straßer, Joke Meheus, and Mathieu Beirlaen. Tolerating deontic conflicts by adaptively restricting inheritance. *Logique et Analyse*, 219:477–506, 2012.

[Straßer *et al.*, 2016] C. Straßer, M. Beirlaen, and F. Van De Putte. Dynamic proof theories for input/output logic. *Studia Logica*, 104:869–916, 2016.

[Straßer *et al.*, 2017] C. Straßer, A. Knoks, and Joke Meheus. Deontic reasoning on the basis of consistency considerations, 2017. Unpublished manuscript, available at `http://www.clps.ugent.be/research/publications`.

[Straßer, 2010] Christian Straßer. An adaptive logic framework for conditional obligations and deontic dilemmas. *Logic and Logical Philosophy*, 19(1-2):95–128, 2010.

[Straßer, 2011] Christian Straßer. A deontic logic framework allowing for factual detachment. *Journal of Applied Logic*, 9:61–80, 2011.

[Straßer, 2014] C. Straßer. *Adaptive Logic and Defeasible Reasoning. Applications in Argumentation, Normative Reasoning and Default Reasoning.* Springer, 2014.

[van Benthem *et al.*, 2014] Johan van Benthem, Davide Grossi, and Fenrong Liu. Priority structures in deontic logic. *Theoria*, 80(2):116–152, 2014.

[van Benthem, 2004] Johan van Benthem. What one may come to know. *Analysis*, 64(282):95–105, 2004.

[Van De Putte and Straßer, 2012] Frederik Van De Putte and Christian Straßer. Extending the standard format of adaptive logics to the prioritized case. *Logique et Analyse*, 220:601–641, 2012.

[Van De Putte and Straßer, 2013] Frederik Van De Putte and Christian Straßer. A logic for prioritized normative reasoning. *Journal of Logic and Computation*, 23(3):563–583, 2013.

[Van De Putte and Straßer, 2014] Frederik Van De Putte and Christian Straßer. Adaptive logics: a parametric approach. *Logic Journal of IGPL*, 22(6):905–932, 2014.

[Van De Putte, 2012] Frederik Van De Putte. *Generic Formats for Prioritized Adaptive Logics. With Applications in Deontic Logic, Abduction and Belief Revision.* Dissertation, Ghent University, 2012. Available at `http://www.clps.ugent.be/research/doctoral-dissertations`.

[Van De Putte, 2013] Frederik Van De Putte. Default assumptions and selection functions: A generic framework for non-monotonic logics. In Felix Castro, Alexander Gelbukh, and Miguel Gonzalez, editors, *Advances in Artificial Intelligence and Its Applications*, volume 8265 of *Lecture Notes in Computer Science*, pages 54–67. Springer, 2013.

[Van De Putte, 2016] Frederik Van De Putte. Coarse Deontic Logic. pages 256–271, July 2016.

[Van De Putte, 2019] Frederik Van De Putte. Coarse deontic logic. *Journal of Logic and Computation*, 29(2):285–317, 2019.

[van der Torre and Villata, 2014] L. van der Torre and S. Villata. An ASPIC-based legal argumentation framework for deontic reasoning. In *Computational Models of Argument (Proceedings of COMMA 14)*, pages 421–432. IOS Press, 2014.

[van Eck, 1982] J.A. van Eck. A system of temporally relative modal and deontic predicate logic and its philosophical applications. *Logique et Analyse*,

99:249–290, 1982.

[van Fraassen, 1973] Bas C. van Fraassen. Values and the heart's command. *Journal of Philosophy*, 70(1):5–19, 1973.

[Verdée, 2013] Peter Verdée. Non-monotonic set theory as a pragmatic foundation of mathematics. *Foundations of Science*, 18(4):655–680, Nov 2013.

[von Wright, 1951] Georg Henrik von Wright. Deontic logic. *Mind*, 60:1–15, 1951.

[Vranas, 2007] P. Vranas. I ought, therefore I can. *Philosophical Studies*, 136:167–216, 2007.

[Williams and Atkinson, 1965] Bernard Williams and W.F̃. Atkinson. Symposium: Ethical consistency. *Proceedings of the Aristotelian Society, Supplementary Volumes*, 39:103–138, 1965.

Frederik Van De Putte
Erasmus University of Rotterdam, The Netherlands;
Ghent University, Belgium
Email: frederik.vandeputte@ugent.be

Mathieu Beirlaen
Ghent University, Belgium

Joke Meheus
Ghent University, Belgium
Email: Joke.Meheus@UGent.be

External Relations

6
Practical Reasoning: Problems and Prospects

Richmond H. Thomason

Abstract. Deontic logic, as traditionally conceived, delivers a logic that supports deduction and that is interpreted in terms of the states or possible worlds within which an agent should seek to remain. As such, it only encompasses a small part of practical reasoning, which is concerned with planning, selecting, committing to, and executing actions. In this chapter I try to frame the general challenge that is presented to logical theory by the problem of formalizing practical reasoning, and to survey the existing resources that might contribute to the development of such a formalization. I conclude that, while a robust, adequate logic of practical reasoning is not yet in place, the materials for developing such a logic are now available.

1 The challenge of formalizing practical reasoning 464
1.1 Some Examples 464
1.2 Towards a classification 472
1.3 Disciplines and approaches 473
1.3.1 Philosophy 473
1.3.2 Logic 476
1.3.3 Psychology 479
1.3.4 Decision Theory and Game Theory 480
1.3.5 Computer science and artificial intelligence 482

2 Towards a formalization 489
2.1 Relaxing the demands of formalization 489
2.2 Agent architectures and division of logical labor 490
2.3 Means-end reasoning 491
2.4 The practicalization of desires 491
2.5 Intention formation 493
2.6 What to do now? 493
2.7 Scheduling, execution and engagement 493
2.8 Framing a practical problem 494

Handbook of Deontic Logic and Normative Systems, Volume 2

1 The challenge of formalizing practical reasoning

Practical reasoning is deliberation. It is reasoning about what to do. We do it all the time. Any day in our life will provide us with hundreds of examples of thinking about what to do. But it has been remarkably difficult to produce a comprehensive, adequate theory of practical reasoning. Part of the difficulty is that the topic is studied by different disciplines, each of these has something important to contribute, and it is unusual to find a study of practical reasoning that brings all of these perspectives together.

This section will begin by considering examples of practical reasoning. Practical reasoning is not like reasoning in mathematics, which is relatively uniform. Even moderately ambitious studies of practical reasoning need to begin with a good picture of the diverse specimens of reasoning that need to be formalized. I will then propose a rationale for classifying these examples, and canvass the disciplines that have something useful to say about the reasoning.

In the remainder of the paper, I try to say something about what a moderately comprehensive logic of practical reasoning might be like.

1.1 Some Examples

All too many published discussions of practical reasoning—even book-length discussions—cover only a very small part of the territory. For that reason, it's vital to begin with a broad range of examples.

Example 1. Ordering a meal at a restaurant.

> The deliberating agent sits down in a restaurant and is offered a menu. Here, the problem is deciding what to eat and drink. Suppose that the only relevant factors are price and preferences about food combinations. Even for a moderately sized menu and wine list, the number of possible combinations is over $400,000$. Naturally, a human decision-maker is not going to compare each of these options explicitly. It may not even be realistic to suppose that that there is a total preference ordering, or that a decision will require a combination that is clearly optimal.
>
> The reasoning in cases like this typically involves quick heuristics for selecting a reasonably good choice, quick identification of alternatives to this choice, and comparison and weighing of tradeoffs. Human agents may sometimes dither on such occasions, but frequently they can easily arrive at a choice with confidence that it is satisfactory.

Example 2. Deciding what move to make in a chess game.

In chess, an individual action needs to be evaluated in the context of its continuations. There is no uncertainty about the current state or the immediate consequences of actions, but much uncertainty about moves that the opponent might make. The *search space* (i.e., the number of possible continuations) is enormous—on the order of 10^{43}. Determining the value of positions involves conflicting criteria (e.g. positional advantages versus numerical strength); these conflicts must be resolved in comparing the value of different positions. In tournament chess, deliberation time is limited. These somewhat artificial constraints combine to concentrate the reasoning on efficient exploration of a search space. Perhaps because of this, the reasoning involved in chess has been intensively investigated by psychologists and computer scientists, and influenced the classical work on search algorithms in AI; see [Simon and Schaeffer, 1992].

Example 3. Savage's omelet.

In [Savage, 1972, pp. 13–15] Leonard Savage describes the problem as follows.

> Your wife has just broken five good eggs into a bowl when you come in and volunteer to finish making the omelet. A sixth egg, which for some reason must either be used for the omelet or wasted altogether, lies beside the bowl. You must decide what to do with the unbroken egg. ... you must decide between three acts only, namely, to break it into the bowl containing the other five, to break it into a saucer for inspection, or to throw it away without inspection.

This problem involves preferences about the desired outcomes. There is risk, too, in the form of a positive probability that the egg is spoiled. The problem is to infer preferences over actions. The outcomes are manifest and involve only a few variables, the preferences over them are evident, and (let's suppose) the probabilities associating each action with an outcome can be easily estimated. In this case, the reasoning task reduces to the calculation of an expected utility.

Example 4. Designing a house.

This example is less obviously practical; it is possible for an architect to design a house without thinking much about the actions that will go into building it, leaving this to the contractor.[1] However, an architect's design becomes the builder's goals, and I would maintain that inferring goals is a form of practical reasoning. The reasoning combines constraint satisfaction and optimization, where again conflicts between competing desiderata may need to be resolved. Any real-life architect will also use *case-based reasoning*, looking in a library of known designs for one that is relevant, and modifying a chosen example to suit the present purpose.

This case is like Example 1, but with added complexity due to multiple constraints. Also, the execution phase may bring to light flaws in the plan, and require changes in the design.

Example 5. Ordering dessert.

Let's return the restaurant of Example 1. The main course is over, and our agent is offered a dessert menu and the choice of whether to order dessert. On the one hand, there is a direct desire for dessert, perhaps even a craving. This alternative is colored with and motivated by emotion, even if the emotion is not overwhelming. But suppose that there is a contrary emotion. The agent is unhappy with being overweight and has determined to eat less, and may even have told others at the table about the decision to undertake a diet. This creates a conflict, coloring the choice of dessert with negative associations, perhaps even shame. The chief difference between this conflict and those in Examples 2 and 4 is that this decision is emotionally "warm;" the outcome may be influenced by a craving and the presence of the desired object. (Perhaps this is why some restaurants invest in dessert trays.)

[1] Of course, a good design has to take into account how to build a house, in order to make sure that the design is feasible.

Example 6. A somewhat shady decision.

The following case description is from the web pages of the National Society of Professional Engineers.

> Engineer P is a top official in the State X highway department. He would like to leave his position to become an executive with an architecture/engineering firm. Engineer P requests permission from the state to accept the new position. State X refuses to grant permission, noting that, in accepting the position, Engineer P would be in violation of the state law that requires top state highway officials to wait a year after leaving the State X highway department before accepting positions with firms with which the department does business. Engineer P leaves the State X highway department and joins the architecture/engineering firm not as an "employee" but as an "independent contractor."[NSPE Board of Ethical Review, 2016]

In deciding that he would like to accept the firm's offer, Engineer P will already have weighed many considerations, including salary, location, job quality, and other factors. The relevant preferences need to be reconsidered when the state refuses permission. Engineer P consults an attorney, learns that he will almost certainly be prosecuted if he accepts the executive position, and judges that the risk of prosecution makes accepting the offer undesirable. Further consultation with the attorney and negotiation with the firm reveals that a temporary consulting position would yield a comparable salary and would be unlikely to result in successful prosecution. Taking into account the new information, Engineer P decides that accepting the consulting position is preferable to continuing with his highway department job.

This case is like Example 5, in that it involves a choice between alternatives. There may be an element of emotional coloring or temptation, if the salary difference is significant. There also is an element of risk—the lawyer may advise that there is a chance of prosecution for accepting the consulting position, but that conviction will turn on the interpretation of 'position' in the statute, and that he would have a strong case if prosecuted. Also, there is an ethical dimension: Engineer P may have taken personal values and ethics into consideration in weighing alterna-

tives, but perhaps he treated these on a par with other factors, such as salary. Clearly, the official didn't treat ethical prohibitions as overriding constraints on the available options.

The NPSE Board of Ethical Review decided the engineer's conduct in this case was unethical.

Example 7. Deciding how to get to the airport.

This is a planning problem; the agent a has an inventory of actions, knows their preconditions and effects, knows the relevant features of the current state, and has as its goal a state in which a is at the airport. In its simplest form, the problem is to find a sequence of actions that will transform the current state into a state that satisfies the goal. Planning, or means-end reasoning, is one of the most intensively studied forms of reasoning in AI. The earliest planning algorithms made many simplifying assumptions about the planning situation and the conditions that a satisfactory plan must meet; over the years, sophisticated planning algorithms have been developed that depend on fewer of these assumptions and so can be used in a variety of realistic settings.[2]

Example 8. Cracking an egg into a bowl.

This is a case in which most of of us do the action automatically, with hardly any conscious reasoning. Probably most people can't remember the circumstances under which they learned how to do it. But the activity is not simple and there are many ways to get it wrong. This example was proposed as a benchmark problem in the formalization of common-sense reasoning. The literature on this problem shows that the reasoning is complex and that it presupposes much common-sense knowledge; see, for instance, [Shanahan, 1997a].

This example is different from the previous ones in that the solution to the reasoning problem is acted out; the reasoning must engage motor systems, and it depends on these systems for grasping and manipulating objects according to plan. For obvious reasons, Savage ignored this part of the omelet problem.

[2]See, for instance, [Reiter, 2001]. For the airport problem in particular, see [Lifschitz *et al.*, 2000].

Example 9. Playing table tennis.

Unlike chess, table tennis is a game in which practical reasoning has to be *online*; engaged in complex, real-time activities involving the perceptual and motor systems. For a novice, the reasoning may be consumed by the need to keep the ball in play; experts may be able to engage in tactical reasoning. But there is no time to spare for reflection; the reasoning needs to be thoroughly connected to the ongoing process of play.

Example 10. Playing soccer.

Soccer is like table tennis, but with the added dimension of teamwork and the need to recognize and execute play. This task was selected as a benchmark problem in robotics, and has been extensively studied. See, for instance, [Visser and Burkhard, 2007, Ros *et al.*, 2009, Asada *et al.*, 1999].

Example 11. Typing a message.

Typing an email message, composing it as you go along, starts perhaps with a general idea of what to say. The reasoning that produced a rough idea of the content may have taken place reflectively, but once composition has begun, several reasoning processes are engaged simultaneously, and have to be coordinated. The general idea of what to say has to be packaged in linguistic form, and this form has to be rendered by motor actions at the keyboard. For a skilled typist composing a straightforward message, these complex, practical tasks are combined and executed very quickly, perhaps at the rate of 70 words per minute. For this to happen, the interface between high-level linguistic reasoning and motor skills has to be very robust.

Example 12. Factory scheduling.

The factory scheduler has to produce, say on a daily basis, a sequence of manufacturing operations for each order to be processed that day, and a schedule allocating times and machines to these operations. This problem is notorious for the difficulty of the reasoning; it involves horrible combinatorics, uncertainty, limited time for reflection, and the resolution of many conflicting desiderata. Among the goals cited by [Fox and Clarke, 1991] are (1) meeting order dates, (2) minimizing work-in-process time, (3) maximizing allocation of factory resources, and (4) minimizing disruption of shop activity.

Part of the interest of this example lies in the difference in scale between this problem and Savage's omelet problem. It is not clear that there is any way to construct a single, coherent utility function for the task, by reconciling the four desiderata mentioned above. Any reconciliation will leave some managers unhappy: salesmen will favor goal (1), and production managers will favor goals (2)-(3), perhaps giving different weights to these. Nor is it easy to produce a global probability function for a system with so many interacting variables.

Example 13. An unanticipated elevator.

A man decides to visit his stockbroker in person, something he has never done. He takes a bus to a stop near the stockbroker's downtown address, gets off the bus, locates the building and enters it. He finds a bank of elevators, and sees that the stockbroker is on the 22nd floor. This man has a strong dislike for elevators, and is not feeling particularly energetic that day. He reconsiders his plan.

Example 14. A woman is working in her garden.

She becomes hot and tired, and decides to take a break. Or she hears the telephone ringing in her house, and decides to answer it. Or she sees smoke coming out of the window of her house, and runs for help.

Example 15. The wrath of Achilles.

In Book I of *The Iliad*, the hero Achilles is outraged and dishonored by his warlord Agamemnon, who insults him and declares that he will take back, in compensation for his own loss and Achilles' disrespectful behavior, the captive woman that Achilles had received as his war prize.

Homer goes on to describe Achilles' reaction. Achilles is headstrong, but his reaction is partly physical and partly intellectual: his heart pounds with rage, but instead of acting immediately he asks himself a question: should he draw his sword and kill the king? To explain his decision, the poet brings in a god: Athena, invisible to everyone else, seizes him by the hair and persuades him to give in and be patient.

For our purposes, we can suppose that Athena is a literary device. The outrage leads to a direct desire to kill, but instead of acting on it, Achilles realizes that it would be better to restrain himself.

Reasoning intervenes here between the strong emotional impulses, and inhibits a reckless action.

Example 16. Deciding what to say at a given point in a conversation.

Conversation provides many good examples of deliberative reasoning. Where there is conscious deliberation, it is likely to be devoted to content selection, and here many factors have to be taken into account: a speaker may need to be interesting, to keep some information private, to address a topic, to be helpful, and to be polite. But once content has been selected, the reasoning that goes into deciding how to express this content can be quite complex.

Certainly, any adequate theory of practical reasoning must at least be compatible with this broad range of cases. Better, it should be capable of saying something about the reasoning involved in all of them. Even better, there should be a single architecture for practical reasoning, capable of dealing with the entire range of reasoning phenomena.[3] No doubt, there are special-purpose cognitive modules (e.g., for managing perception, motor behavior, and some aspects of language). But it would be perverse to formulate a theory of a special type of practical reasoning, such as preference generation, probability estimation, or

[3]For the idea of a cognitive architecture, see [Newell, 1992].

means-end reasoning, and to postulate a specialized module that performs just this reasoning. The interaction between these components of practical reasoning is too strong for modularization to be feasible, and this methodology would be likely to produce an *ad hoc* and piecemeal account of practical reasoning.

1.2 Towards a classification

The examples in the previous section suggest a set of features that can be used to classify specimens of deliberative reasoning.

1. Are only a few variables (e.g., desiderata, causal factors, initial conditions) involved in the decision?
2. Do conflicting preferences need to be resolved in making the decision?
3. Is the time available for deliberation small compared to the time needed for adequate reflection?
4. Is the deliberation immediate? That is, will the intentions that result from the deliberation be carried out immediately, or postponed for future execution?
5. Is the deliberation carried out in "real time" as part of an ongoing activity involving sensory and motor activities?
6. Does the reasoning have to interface closely with sensory and motor systems?
7. Is the activity part of a group or team?
8. Does the context provide a definite, relatively small set of actions, or is the set of actions open-ended?
9. Is there certainty about the objective factors that bear on the decision?
10. Is the associated risk small or great?
11. Is the goal of deliberation a single action, or a sequence of actions?
12. Is continuous time involved?
13. Is the deliberation colored with emotions?

14. Is the action habitual, or automatic and unreflective?
15. Is there conscious deliberation?
16. Are there existing plans in play to which the agent is committed or that already are being executed?

Many of the differences marked by these features are matters of degree, so that the boundaries between the types of reasoning that they demarcate are fluid. This strengthens the case for a general approach to the reasoning. There is nothing wrong with concentrating on a special case to see what can be learned from it. Chess and decision problems that, like Savage's omelet, take advantage of a solution to the "small worlds problem"[4] provide good examples of cases where this methodology has paid off. But to concentrate on these cases without paying any attention to the broad spectrum of examples runs the risk of producing a theory that will not be contribute usefully to something more general.

1.3 Disciplines and approaches

Many different disciplines have something to say about practical reasoning. The main theoretical approaches belong to one of the five following areas.

1. Philosophy
2. Logic
3. Psychology
4. Decision theory and game theory
5. Artificial intelligence

Of course, there is a good deal of overlap and mixing of these approaches: AI, for instance, is especially eclectic and has borrowed heavily from each of the other fields. But work in each area is influenced by the typical problems and methods of the discipline, and—typically, at least—has a distinctive perspective that is inherited from the parent discipline.

The following discussion of these five approaches is primarily interested in what each has to contribute to the prospects for formalizing practical reasoning.

1.3.1 Philosophy

The topic of practical reasoning goes back to Aristotle. In the twentieth century there was a brief revival of philosophical interest in "practical

[4]This is the problem of framing a decision problem, by isolating and quantifying the relevant factors.

inference." This coincided more or less with early work on deontic and imperative logic, and was carried out by a group of logically minded philosophers and a smaller group of philosophically minded logicians. It is a little difficult to distinguish philosophy from logic in this work; I will more or less arbitrarily classify Kenny and some others as philosophers for the purposes of this exposition, and von Wright as a logician.

Post-Fregean interest in imperative logic seems to have begun about the time of World War 2, with [Jørgensen, 1937-1938, Hofstadter and McKinsey, 1939, Ross, 1941]. Later, in the 1960s,[5] some British philosophers became interested in the topic. This period saw 10 or more relevant articles appearing in journals like *Analysis*. Of these, [Kenny, 1966] seems to have the most interesting things to say about the problem of formalizing practical reasoning.[6]

Kenny begins with Aristotle's practical syllogism, taking several specimens of means-end reasoning from the Aristotelian corpus, and beginning with the following example, based on a passage in *Metaphysics* 1032b19.

Example 17. A doctor prescribing.

This man is to be healed.
If his humors are balanced, he will be healed.
If he is heated, his humors will be balanced.
If he is rubbed, he will be heated.
So I'll rub him.

The premisses of the reasoning, according to Kenny, are either (i) desires or duties, or (ii) relevant facts. And he characterizes the conclusion as an action.[7] Kenny points out that this sort of reasoning doesn't fit Aristotelian syllogistic, and that a straightforward modern formalization of it would be invalid. To put it crudely, the inference from P, $Q \to P$, $R \to Q$, and $S \to R$ to S is invalid.

[5] Judging from internal evidence, the work of Richard Hare influenced this episode of interest in the topic. Elizabeth Anscombe [Anscombe, 1958] may also have been an influence, as well as G.H. von Wright.

[6] For more about this period, see [Green, 1997].

[7] The Aristotelian texts make it pretty clear that Aristotle considered the conclusion to be an action. But for our purposes, it would work better to think of the conclusion as an expression of intention. In some circumstances—when the deliberation is concerned with immediate action and the reasoning is sufficiently persuasive, there is no gap between intention and action.

Here, Kenny has indicated an important type of practical reasoning, and pointed out a glaring problem with the propositional calculus as a formalization medium. Unfortunately, the theory that he proposes in this paper doesn't seem to solve the problem of providing an account of validity that matches the reasoning.

In fact, there are many glaring problems with the crude Propositional Calculus formalization of Example 17, involving the deductive formulation of the reasoning as well as the faithfulness of the formalization to the language of the example. The failure of Kenny's proposal and of similar ones at the time seems to originate in a lack of logical resources that do justice to the problem. The Propositional Calculus is certainly not the right tool, and deduction is certainly not the right characterization of the reasoning. The only idea that was explored at the time was that of providing a logic of "imperative inference." This idea might help with one problem: formalizing the first premiss of Example 17, which does not seem like a straightforward declarative. But it can't begin to address the challenge posed by the invalidity of the argument. Besides, the idea of an imperative logic didn't lead to anything very new, because of another trend that was taking place at about the same time.

This trend, which tried to absorb imperative and practical inference into some sort of modal logic, was also underway in the 1960s. [Lemmon, 1965] provides a logic of imperatives that prefigures the STIT approach of [Belnap, Jr. *et al.*, 2001]: a modal approach that brings in the idea of causing a state of affairs. And [Chellas, 1969], recommends and develops a reduction of imperative logic to a more standard deontic logic. This idea provides formal systems with excellent logical properties. But it does so at the expense of changing the subject, and leaving the central problem unsolved. Reasoning in deontic logic is deductive, and if you formalize typical specimens of means-end reasoning like Example 17 in these systems, the formalizations will be invalid.

Even though the literature shows a sustained series of attempts in this period to formalize practical inference, the work didn't lead to anything like a consensus, and produced no sustainable line of logical development. In retrospect, we can see that the formalization project was unsustainable because of oversimplification, and in particular because of the assumption that a successful formalization must be deductive and resemble propositional truth-functional logic.

As we will see in Section 1.3.5, more recent and quite separate developments in computer science have yielded sophisticated logics of means-end reasoning, effectively solving the formalization problem that led to an earlier philosophical impasse in the 1960's and 1970's. The moral

seems to be that formalization projects of this sort can involve multiple challenges, and that it can be hard to address these challenges without a body of applications and a community of logicians committed to formalizing the applications and mechanizing the reasoning.

Meanwhile, philosophers seem to have drawn the conclusion that close attention to the reasoning and attempting to formalize it is not likely to be productive. In the more recent philosophical work on practical reasoning, it is actually quite difficult to find anything that bears on the formalization problem. Almost entirely, the philosophical literature is devoted to topics that might serve to provide philosophical foundations for the theory of practical reasoning—if there were such a theory. Even if, as Elijah Millgram claims in [Millgram, 2001], the driving issue in the philosophy of practical reasoning is to determine which forms of practical reasoning are correct, philosophers seem to pursue this inquiry with informal and hazy ideas of the reasoning itself. In many cases—for instance, the issue of whether intentions cause actions—no formalization of the reasoning is needed for the philosophical purposes. In other cases, however, a formal theory of practical reasoning might help the philosophy, refining some old issues and suggesting new ones. I myself would go further, and say that it is premature and pointless to philosophize about the foundations of a theory before the theory is in place.

Even though some philosophers maintain positions that would sharply limit the scope of practical reasoning (reducing it, for instance, to means-end reasoning), I don't know of any explicit, sustained attempt in the philosophical literature to delineate what the scope of practical reasoning should be. I don't see how to do this without considering a broad range of examples, as I try to do above in Section 1.1. But in fact, examples of practical reasoning are thin on the ground in the philosophical literature; in [Millgram, 2001], for instance, I counted only 12 examples of practical reasoning in 479 pages—and many of these were skeleton examples, intended as illustrations of general points.

1.3.2 Logic

Contemporary logic is departmentalized. Work in logic bearing on practical reasoning tends to be carried out in the context of either philosophy or computer science, and to be influenced by the interests of the parent disciplines. There are, in fact, two separate strands of logical research, one associated with philosophy and the other with artificial intelligence. These have interacted less than one might wish. I'll discuss the philosophical tradition first, and will conclude this section with the later,

computational tradition.

Georg Henrik von Wright was explicitly interested in practical reasoning, from both a philosophical and a logical standpoint. Most of his writings on the topic are collected together in [von Wright, 1983]; these were published between 1963 and 1982. Like Kenny, von Wright begins with Aristotle's practical syllogism. But he avoids the problem of invalidity by strengthening the premisses. Aristotle's formulations involve a premiss to the effect that a course of action is *a way* of achieving a goal; Von Wright changes this to the stronger claim that it is *the only* way. For instance, von Wright's version of Example 17 would look like this:

> I want to heal this man.
> Unless his humors are balanced, he will not be healed.
> Unless he is heated, his humors will not be balanced.
> Unless he is rubbed, he will not be heated.
> Therefore I must rub him.

By departing from Aristotle's formulation, von Wright makes it easier to formulate the inference in a deontic logic, and to see how the formalization might be valid. But there is a price for this; in general, von Wright's premisses will be implausible. In the medical example, for instance, there surely will be more than one way to heat the patient and the deliberating agent must choose among the many ways to do this. In fact, choosing between alternative means is characteristic of means-end reasoning, and to ignore this is to miss something important.

Von Wright's simplification makes it easier for him to propose modal logic, and in particular deontic logic, as the formalization medium for practical reasoning. Von Wright also characterizes his version of deontic logic as a "logic of action." All this seems to mean is that the atomic formulas of his language may stand for items of the form 'Agent A does action a.' But a practical agent with a goal in mind chooses an action because its consequences will help to realize the goal — and realization belongs to the causal order of things. So a formalization of practical reasoning must tell us what the consequences of actions are, and how they enter into the causal order. In this respect, Von Wright does not provide a logic of action.

I will not say much here about the subsequent history of deontic logic as a part of philosophical logic. As the field developed, it acquired its own problems and issues (such as the problem of reparational obligations), but as philosophers concentrated on declarative formalisms and deductive logic, the relevance to practical reasoning, and even means-end reasoning, that von Wright hoped for in his in early papers, such as

[von Wright, 1963], became attenuated.

Although the subsequent history of deontic logic was less directly concerned with practical reasoning, it shows a healthy tendency to concentrate on naturally occurring problems that arise in reasoning about obligation. This work, which certainly will be well documented in the present volume, has a place in any general theory of practical reasoning. Obligations play a role as constraints on means-end reasoning, and reasoning about obligations must be flexible to cope with changing circumstances.

Also, the problem of modeling conditional obligations has produced a large literature on the relationship between modal logic and preference.[8] Of course, reasoning about preferences intrudes into practical reasoning in many ways. How to fit it in is something I am not very clear about at the moment; part of the problem is that so many different fields study preferences, and preferences crop up in so many different types of practical reasoning. Maybe the best thing would be to incorporate preferences in a piecemeal way, and hope that a more general and coherent approach might emerge from the pieces.

The STIT approach to agency was already mentioned in Section 1.3.1. This provides a model-theoretic account of how actions are related to consequences that is quite different from the ones that emerged from the attempts in AI to formalize planning. The connections of this theory to practical reasoning are tenuous, and I will not have much to say about STIT.

Philosophy and philosophical logic have served over the years as a source of ideas for extending the applications of logic, and for developing logics that are appropriate for the extensions. One would hope that philosophy would continue to play this role. But—at least, for areas of logic bearing on practical reasoning—the momentum has shifted to computer science, and especially to logicist AI and knowledge representation. This trend began around 1980, and has accelerated since. Because many talented logicians were attracted to computer science, and because the need to relate theories to working implementations provided motivation and guidance of a new kind, this change of venue was accompanied by dramatic logical developments, and improved insights into how logic fits into the broader picture. I would very much like to see philosophy continue to play its foundational and creative role in developing new applications of logic, but I don't see how this can happen in the area of practical reasoning unless philosophers study and assimilate the recent contributions

[8] See, for instance, [Hansson, 2001, Jones and Carmo, 2002].

of computer scientists.

The point is illustrated by [Gabbay and Woods, 2005]. The paper is rare among contemporary papers in urging the potential importance of a logic of practical reasoning, but—in over 100 pages—it is unable to say what a coherent, sustained research program on the topic might be like. It does mention some important ideas, such as taking the agent into account, as well as nonmonotonic and abductive reasoning, but offers no explicit, articulated theories and in fact is hesitant as to whether logic has a useful role to play, repeating some doubts on this point that have been expressed by some roboticists and cognitive psychologists. Although it cites a few papers from the AI literature, the citations are incidental; work on agent architectures, abductive reasoning, and means-end reasoning goes unnoticed. Part of the problem is that the authors seem to feel that work in "informal logic" might be useful in approaching the problem of practical reasoning—but the ideas of informal logic are too weak to provide any helpful guidance. If we are interested in accounting for the practical reasoning of agents, computer programs must play a part in our accounts. For this, we need formal logic—but formal logic that is applicable.

I couldn't agree more with Gabbay and Woods that logicians should be concerned with practical reasoning. But to make progress in this area, we need to build on the accomplishments of the formal AI community.

1.3.3 Psychology

Early practitioners of cognitive psychology invested considerable effort into the collection of protocols from subjects directly engaged in problem-solving, much of it practical. Herbert Simon and Allen Newell were early and peristent advocates of this methodology. Their protocols contain many useful examples; in fact, they helped to inspire early characterizations of means-end reasoning in artificial intelligence.

As early as 1947, in [Simon, 1947], Simon had noted divergences between decision-making in organizations and the demands of ideal rationality that are incorporated in decision theory; he elaborated the point in later work. An important later trend that began in psychology, with the work of Amos Tversky and Daniel Kahneman, studies these differences in more detail, providing many generalizations about the way people in fact make decisions and proposing some theoretical models; see, for instance, [Kahneman and Tversky, 1979, Tversky and Kahneman, 1981].

Tversky and Kahneman's experimental results turned up divergences

between ideal and actual choice-making that were not obviously due, as Simon had suggested, merely to the application of limited cognitive resources to complex, time-constrained problems. Since their pioneering work, this has become a theme in later research.

All this raises a challenging foundational problem, one that philosophers might be able to help with, if they gave it serious attention. What level of idealization is appropriate in a theory of deliberation? What is the role of "rationality" in this sort of idealization? Is there a unique sort of rationality for all practically deliberating agents, or are there many equally reasonable ways of deliberation, depending on the cognitive organization and deliberative style of the agent? Is the notion of rationality of any use at all, outside the range of a very limited and highly idealized set of decision problems? Probably it would be unwise to insist on solutions to these problems before attempting to provide a more adequate formalization of practical reasoning—that would be likely to delay work on the formalization indefinitely. But the problems are there.

Nowadays, the cognitive psychology of decision-making has migrated into economics and management science, and is more likely to be found in economics departments and business schools than in psychology departments. This doesn't affect the research methods much, but it does improve the lines of communication between researchers in behavioral economics and core areas of economics. As a result, economic theorists are becoming more willing to entertain alternatives to the traditional theories.

1.3.4 Decision Theory and Game Theory

The literature in these areas, of course is enormous, and most of it has to do with practical reasoning. But traditional work in game theory and decision theory concentrates on problems that can be formulated in an idealized form—a form in which the reasoning can be reduced to the calculation of an optimum result.[9] As a result, work in this

[9]Microeconomists and statisticians are not the only ones who have taken this quantitative, calculational paradigm to heart. Many philosophers have accepted the paradigm as a model of practical reasoning and rationality. See, for instance, [Skyrms, 1990], a book-length study of practical deliberation, which supposes the only relevant theoretical paradigms to be decision theory and game theory, and takes them pretty much in the classical form. Skyrms' book and the many other philosophical studies along these lines have useful things to say; my only problem with this literature is the pervasive assumption that practical reasoning can be comprehensively explained by quantitative theories which presuppose that global probability and utility functions

tradition tends to neglect much of the reasoning in everyday practical reasoning, which seldom involves quantitative estimates of probabilities and utilities, explicit calculation, or insistance on an optimum.

Of course, an agent must reason to wrestle a practical problem into the required form—Savage's small worlds problem is a reasoning problem, with better and worse solutions. But the literature in economics tends to assume that somehow the problem has been framed, without saying much if anything about the reasoning that might have gone into this process. (Work in decision analysis, of course, is the exception.) And once a problem has been stated in a form that can be solved by calculation, there is little point in talking about deliberative processes.

If we are concerned with the entire range of examples presented in Section 1.1, however, we find many naturally occurring problems that must be addressed without a solution to the small worlds problem. People manage to deliberate successfully in such cases without formulating their problems in decision theoretic terms. And their decisions are based on reasoning that is often quite complex. This is one reason why I believe that there will be an important place in a general theory of practical reasoning for qualitative reasoning, and especially for inferential reasoning—the sort of reasoning that gives formalization and logic a foothold. In this respect, Aristotle was on the right track.

At the very least, practical reasoning can involve inference and heuristic search, as well as calculation. (Calculation, of course, is a form of reasoning, but is not inferential, in the sense that I intend.) Any theory of practical reasoning that emphasizes one sort of reasoning at the expense of others must sacrifice generality, confining itself to only a small part of the territory that needs to be covered by an adequate approach. The imperialism of some of those (mainly philosophers, these days) who believe that there is nothing to rationality or practical reasoning other than calculations involving probability and utility, can partly be excused by the scarcity of theoretical alternatives. I will argue in this chapter that the field of artificial intelligence has provided the materials for developing such alternatives.

As I said in Section 1.3.3, research in behavioral economics has made microeconomists generally aware that, in their original and extreme form, the idealizations of decision theory don't account well for a broad range of naturally occuring instances of practical reasoning. Attempts to mechanize decision making led computer scientists to much the same conclusion.

can be associated with an agent.

A natural way to address this problem begins with decision theory in its classical form and attempts to relax the idealizations. Herbert Simon made some early suggestions along these lines; other, quite different, proposals can be found in [Weirich, 2004] and [Russell and Wefald, 1991]. And other relaxations of decision theory have emerged in artificial intelligence: see the discussion of Conditional Preference Nets below, in Section 1.3.5. Still other relaxations have emerged out of behavioral economics, such as Tversky and Kahneman's Prospect Theory; see [Kahneman and Tversky, 1979].

Programs of this sort are perfectly compatible with what I will propose here. A general account of practical reasoning has to include calculations that somehow combine probability (represented somehow) and utility (represented somehow), in order to estimate risk. The more adaptable these methods of calculation are to a broad range of realistic cases, the better. I do want to insist, however, that projects along these lines can only be part of the story. Anyone who has monitored their own decision making must be aware that not all practical reasoning is a matter of numerical calculation; some of it is discursive and inferential. A theory that does justice to practical reasoning has to include both forms of reasoning. From this point of view, the trends from within economics that aim at practicalizing game theory and decision theory are good news. From another direction, work in artificial intelligence that seeks to incorporate decision theory and game theory into means-end reasoning is equally good news.[10]

In many cases of practical reasoning, conflicts need to be identified and removed or resolved. Work by economists on value tradeoffs is relevant and useful here; the classical reference is [Keeney and Raiffa, 1976], which contains analyses of many naturally occuring examples.

1.3.5 Computer science and artificial intelligence

For most of its existence, the field of AI has been concerned with realistic decision problems, and compelled to formalize them. As the field matured, the AI community looked beyond procedural formalizations in the form of programs to declarative formulations and logical theories. Often AI researchers have had to create their own logics for this purpose.[11] Here, I will be concerned with three trends in this work: those

[10]For a survey, now getting rather old, see [Blythe, 1999]. For an example of a more recent, more technical paper, see [Sanner and Boutilier, 2009].

[11]Throughout his career, John McCarthy was a strong advocate of this approach, and did much of the most important work himself. See [McCarthy and Hayes, 1969]

that I think have most to offer to the formalization of practical reasoning. These are means-end reasoning, reasoning about preferences, and agent architectures.

Dynamic logic and imperative inference. When an agent is given instructions and intends to carry them out unquestioningly, there is still reasoning to be done, and the reasoning is practical[12]—although, as the instructions become more explicit, the less scope there is for interesting reasoning from the human standpoint. The case of computer programs, where explicitness has to be carried out ruthlessly, can be instructive, because it shows how logical theory can be useful, even when the reasoning paradigm is not deductive.

A computer program is a (possibly very large and complex) imperative — it is a detailed instruction for carrying out a task. Many of its components, such as

$$\textbf{let } \boldsymbol{y} \textbf{ be } \boldsymbol{x}$$

("set the value of y to the current value of x") are imperatives, although some components, like the antecedent of the conditional instruction

$$\textbf{if } (\boldsymbol{x} < \boldsymbol{y} \textbf{ and } \textbf{not}(\boldsymbol{x} = \boldsymbol{0})) \textbf{ then let } \boldsymbol{z} \textbf{ be } \boldsymbol{y}/\boldsymbol{x}$$

are declarative.

Inference, in the form of proofs or a model theoretic logical consequence relation, plays a small part in the theory of dynamic logic. Instead, *execution* is crucial: the series of states that the agent (an idealized computer) goes though when, starting in a given initial state, it executes a program. Because states can be identified with assignments to variables, there are close connections to the familiar semantics of first-order logic.

Dynamic logic is useful because of its connection to *program verification*. A program specification is a condition on what state the agent will reach if it executes the program; if the initial state of a parsing program for an English grammar G, for instance, describes a string of English words, the program execution should eventually halt. Furthermore, (1) if the string is grammatical according to G, the executor should reach a final state that describes a parse of the string, and (2) if the string is not grammatical according to G it should reach a final state that records its ungrammaticality.

for an early statement of the methodology, and a highly influential proposal about how to formalize means-end reasoning.

[12] See [Lewis, 1979].

Dynamic logic has led to useful applications and has made important and influential contributions to logical theory. It is instructive to compare this to the relatively sterile philosophical debate concerning "imperative inference" that took place in the 1960s and early 1970s.[13] To a certain extent, the interests of the philosophers who debated imperative inference and the logicians who developed dynamic logic were different. Among other things, the philosophers were interested in applications to metaethics, and computational applications and examples didn't occur to them.

But the differences between philosophers and theoretical computer scientists are relatively unimportant; some of the philosophers involved in the earlier debate were good logicians, and would have recognized a worthwhile logical project if it had occurred to them. In retrospect, three factors seem to have rendered the earlier debate unproductive:

(1) Too great a reliance on deductive paradigms of reasoning;
(2) Leaving a model of the executing agent out of the theoretical picture;
(3) Confining attention to simple examples.

In dynamic logic, the crucial semantic notion is the correctness of an imperative with respect to a specification. Logically interesting examples of correctness are not likely to present themselves without a formalized language that allows complex imperatives to be constructed, and without examples of imperatives that are more complicated than 'Close the door'. (The first example that is presented in [Harel *et al.*, 2000] is a program for computing the greatest common divisor of two integers; the program uses a *while*-loop.) And, of course, a model of the executing agent is essential to the logical theory. In fact, what is surprising is how much logic can be accomplished with such a simple and logically conservative agent model.

As I said, the activity of interpreting and slavishly executing totally explicit instructions is a pretty trivial form of practical reasoning. But a logic of this activity is at least a start. I want to suggest that, in seeking to formalize practical reasoning, we should be mindful of these reasons for the success of dynamic logic, seeking to preserve and develop them as we investigate more complex forms of practical reasoning.

Planning and the formalization of means-end reasoning. Perhaps the most important contribution of AI to practical reasoning is

[13] See, for instance, [Williams, 1963, Geach, 1963] as well as [Kenny, 1966], which was discussed above, in Section 1.3.1.

the formalization of means-end reasoning, along with appropriate logics. This forms an impressive body of research into the metamathematical properties of these logics, and implementations in planning systems.[14]

This approach to means-ends reasoning sees a planning problem as consisting of the following components:

(1) An initial state. (This might be described by a set of literals—of positive and negative atomic formulas.)

(2) Desiderata or goals. (These might consist of a set of formulas with one free variable; a state that satisfies these formulas is a goal state.)

(3) A set of actions or operators. Each action a is associated with a causal axiom, saying that if a state s satisfies certain preconditions, then a state $\textsc{Result}(a, s)$ that results from performing a in s will satisfy certain postconditions.

Here, the fundamental logical problem is how to define the successor state or the set of these states[15] resulting from the performance of an action in a state. (Clearly, not all states satisfying the postconditions of the action will qualify, since many truths will carry over to the result by "causal inertia.") This large and challenging problem spawned a number of subproblems, of which the best-known (and most widely misunderstood) is the *frame problem*. Although no single theory has emerged from years of work on this problem as a clear winner, the ones that have survived are highly sophisticated formalisms that not only give intuitively correct results over a wide range of test cases, but provide useful insights into reasoning about actions. Especially when generalized to take into account more realistic circumstances, such as uncertainty about the current state and concurrency or nondeterminism, these planning formalisms deliver logical treatments of means-end reasoning that go quite far towards solving the formalization problem for this part of practical reasoning.

I will try to say more about how these developments might contribute to the general problem of formalizing practical reasoning below, in Section 2.3.

Reasoning about preferences It is hard to find AI applications that don't involve making choices. In many cases, it's important to align

[14] [Allen *et al.*, 1990] is a collection of early papers in the field. Both [Shanahan, 1997b] and [Reiter, 2001] describe the earlier logical frameworks and their later generalizations; [Reiter, 2001] also discusses implementation issues.

[15] Depending on whether we are working with the deterministic or the nondeterministic case.

these choices with the designer's or a user's preferences. Implementing such preference-informed choices requires (i) a representation framework for preferences, (ii) an elicitation method that yields a rich enough body of preferences to guide the choices that need to be made, and (iii) a way of incorporating the preferences into the original algorithm.

Any attempt to extract the utilities needed for even a moderately complex, realistic decision problem will provide motives for relaxing the classical economic models of utility; but the need for workable algorithms seems to sharpen these motives. See [Goldsmith and Junker, 2008] for examples and details, and [Doyle, 2004], which provides a wide-ranging foundational discussion of the issues, with many references to the economics literature.

Of the relaxations of preference that have emerged in AI, *Ceteris Paribus* Preference Nets are one of the most widely used formalisms.[16] As in multi-attribute utility theory, the outcomes to be evaluated are characterized by a set of features. A parent-child relation over features must be elicited from a human subject; this produces a graph called a *CP-net*. The parents of a child feature are the features that directly influence preferences about the child. For instance, the price of wheat in the fall (high or low) might influence a farmer's preferences about whether to plant wheat in the spring. If the price will be high, the farmer prefers to plant wheat; otherwise, he prefers not to plant it. On the other hand, suppose that in the farmer's CP-net the price of lumber is unrelated to planting wheat. It can then be assumed that preferences about planting wheat are independent of the price of lumber.

To complete the CP-net, a preference ranking over the values of a child feature must be elicited for each assignment of values to each of the parent features.

Acyclic CP-nets support a variety of reasoning applications (including optimization), and—combined with means-end reasoning—provide an approach to preference-based planning.[17] And in many realistic cases it is possible to extract the information needed to construct a CP-net.

There are extensions of this formalism that allow for a limited amount of reasoning about the priorities of features in determining overall preferences; see [Brafman *et al.*, 2006].

Like decision analysis, work in AI on preferences tends to concentrate on extracting preferences from a user or customer. Thinking about practical reasoning, however, produces a different emphasis. Some of the

[16] See, for instance, [Domschlak, 2002, Boutilier *et al.*, 2003].

[17] See [Baier and McIlraith, 2008] for details and further references.

examples in Section 1.1—for instance, Examples 1, 4, 5, and 12—were designed to show that preferences are not automatically and trivially produced by the environment, by other agents, by the emotions, or by a combination of these things. We can deliberate about what is better than what, and preferences can be the outcome of practical reasoning.[18] The status of an agent trying to work out its own all-things-considered preferences, and of a systems designer or decision analyst trying to work out the preferences of a person or an organization, may be similar in some ways, but I don't think we can expect them to be entirely the same. Nevertheless, insights into methods for extracting preferences from others might be helpful in thinking about how we extract our own preferences.

Agent architectures. A nonexecuting planning agent is given high-level goals by a user, as well as declarative information about actions and the current state of things, as well perhaps as preferences to be applied to the planning process. With this information, it performs means-end reasoning and passes the result along to the user in the form of a plan.

This agent is not so different from the simple instruction-following agent postulated by dynamic logic; its capabilities are limited to the execution of a planning program, and it has little or no autonomy. But—especially in time-limited planning tasks—it may be difficult to formulate a specification, because the notion of what counts as an optimal plan in these conditions is unclear.

When the planning agent is equipped with means of gathering its own information, perhaps by means of sensors, and is capable of performing its own actions, the situation is still more complicated, and is more interesting. Now the agent is interacting directly with its environment, and not only produces a plan, but must adopt it and put it into action. This has a number of important consequences. The agent will need to perform a variety of cognitive functions, and to interleave cognitive performances with actions and experiences.

[18] For some preliminary and sketchy thoughts about this, see [Thomason, 2002].

(1) Many of the agent's original goals may be conditional, and these goals may be activated by new information received from sensors. This is not full autonomy, but it does provide for new goals that do not come from a second party.
(2) Some of these new goals may be urgent; so the agent will need to be interruptable.
(3) It must commit to plans—that is, it must form intentions. These intentions will constrain subsequent means-end reasoning, since conflicts between its intentions and new plans will need to be identified and eliminated.
(4) It will need to schedule the plans to which it has committed.
(5) It will need to monitor the execution of its plans, to identify flaws and obstacles, and repair them.

Recognizing such needs, some members of the AI community turned their attention from inactive planners to *agent architectures*, capable of integrating some of these functions. Early and influential work on agent architectures was presented in [Bratman *et al.*, 1988]; this work stressed the importance of intentions, and the role that they play in constraining future planning.

Any means-end reasoner needs desires (in the form of goals) and beliefs (about the state of the world and the consequences of actions). As Bratman, Israel, and Pollock point out, an agent that is implementing its own plans also needs to have intentions. Because of the importance of these three attitudes in the work that was influenced by these ideas, architectures of this sort are often known as *BDI architectures*. For an extended discussion of BDI architectures, with references to the literature up to 2000, see [Wooldridge, 2000]. See also [Georgeff *et al.*, 1999].

Work in "cognitive robotics" provides a closely related, but somewhat different approach to agent architectures. Ray Reiter, a leading figure in this area, developed methods for integrating logical analysis with a high-level, programming language called GoLog, an extension of Prolog. Reiter's work is continued by the Cognitive Robotics Group at the University of Toronto.

Developments in philosophical logic and formal semantics have provided logics and models for propositional attitudes; for instance, see [Fagin *et al.*, 1995, Fitting, 2009]. Using these techniques, it is possible to formulate a metatheory for BDI agency. Such a metatheory is not an architecture; the reasoning modules of a BDI agent and overall control of reasoning still have to be described procedurally. But the metatheory can provide specifications for some of the important reasoning tasks. Wooldridge's logic of rational agents, $\mathcal{LORA}$, develops this idea; see

[Wooldridge, 2000].

A final word. Logicist AI has struggled to maintain a useful relation to applications, in the form of workable technology. Although the struggle has been difficult, many impressive success stories have emerged from this work—enough to convince the larger AI community of the potential value of this approach. The incentive to develop working applications has, I believe, been very helpful for logic, enabling new ideas that would not have been possible without the challenges posed by complex, realistic reasoning tasks.

Practical reasoning is not quite the same as logicist AI, or even the logical theory of BDI agents. But the successful use of logical techniques in this area of AI provides encouragement for a logical approach to practical reasoning. And, of course, it provides a model for how to proceed.

2 Towards a formalization

The challenge is this: how to bring logical techniques to bear on practical reasoning, and how to do this in a way that is illuminating, explanatory, and useful? In this chapter, I will only try to provide an agenda for addressing this challenge. The agenda divides naturally into subprojects. Some of these subprojects can draw on existing work, and especially on work in AI, and we can think of them as well underway or even almost completed. Others are hardly begun.

2.1 Relaxing the demands of formalization

Let's return to the division between theoretical and practical reasoning.

Traditionally, theoretical reasoning domains are formalized using what Alonzo Church called the "logistic method."[19] This method aims to formulate a formal language with an explicit syntax, a model-theoretically characterized consequence relation, and perhaps a proof procedure. Traditional formalizations did not include a model of the reasoning agent, except perhaps, in the highly abstract form of a Turing machine—this sort of agent is guaranteed whenever the consequence relation is recursively enumerable.

When it comes to practical reasoning, I believe that we have to be prepared to relax Church's picture of logical method.[20] My own proposal

[19] [Church, 1959][pp. 47–58].

[20] In fact, writing in 1956, Church was uncomfortable with semantics and model

for a relaxation is this: (1) we need to add a model of the reasoning agent, (2) we need to identify different phases of practical reasoning in agent deliberation, and different ways in which logic might be involved in each phase of the reasoning, and (3) consequently, we need to be prepared to have a logical treatment that is more pluralistic and less unified.

2.2 Agent architectures and division of logical labor

How should we model an agent that is faced with practical reasoning problems? In Section 1.1, I suggested that we should aim at, or at least acknowledge the existence of, a very broad range of reasoning problems. Suppose, for instance, that we classify the types of reasoning that we may need to consider in terms of the sort of conclusion that is reached. In view of the examples that were presented in Section 1.1, we will need to be prepared for the agent to infer:

(1) Goals, which then invoke planning processes;
(2) Plans, and the subgoals or means that emerge from plans;
(3) Preferences emerging from reasoning about tradeoffs and risk;
(4) Intentions, commitments about what to do, and (to an extent) about when to do it;
(5) Immediate decisions about what plan to execute;
(6) Immediate, engaged adjustments of ongoing activities and plan executions, and shifts of attention that can affect the task at hand.

The examples in Section 1.1 were chosen, in part, to illustrate these activities. These sorts of deliberation are distinct, and all are practical. Although some of them can be automatic, they all can involve deliberate reasoning.

These six activities comprise my (provisional) division of practical reasoning into subtasks, and of the deliberating agent into subsystems. Each of them provides opportunities for logical analysis and formalization. I will discuss them in turn.

theory. He included these topics, but in a whisper, using small type. Over 50 years later, we have become quite comfortable with model theory and semantics, and are more likely to insist on this ingredient than on proof procedures. And in areas where logic is applied, we have become increasingly comfortable with the idea of bringing the reasoning agent into the picture.

2.3 Means-end reasoning

This is the best developed of the six areas. We can refer to the extensive AI literature on planning and means-end reasoning not only for well developed logical theories, but for ideas about how this deliberative function interacts with the products of other deliberative subsystems—for instance, with preferences, and with plan monitoring and execution.

2.4 The practicalization of desires

On the other hand, work in AI on means-end reasoning, and on BDI agents, has little or nothing to say about the emotions and the origins of desires. In general, it is assumed that these come from a user—although the goals may be conditional, so that they are only activated in the appropriate circumstances. In principle, there is no reason why goals couldn't be inferred or learned. But the relevant reasoning processes have not, as far as I know, been formalized.

In truly autonomous agents some desires—perhaps all—originate in the emotions. Although a great deal has been written about the emotions, it is hard to find work that could fit usefully into a logical agenda.[21]

Set aside the issue of how desires originate, and consider only the results of the process. Although the things that are desired are warmed by emotion, they are warmed to a different degree. And, in attraction-avoidance conflicts, they can be warmed and cooled at the same time. To be useful in reasoning, some desires must be conditional, and self-knowledge about conditional desires must be robust. My preference for white wine this evening will probably be colored by feelings of pleasure when I think about the refreshing taste of white wine. But the feeling of hypothetical pleasure is relatively mild; I am certainly not carried away by the feeling. And AI systems builders are interested in obtaining a large body of conditional preferences from users because preferences need to be brought to bear under many different circumstances, so that a user's unconditional preferences—the preferences that are activated in the actual state of affairs—will not in themselves be very useful. Fully autonomous agents need conditional preferences as well, in planning future actions and in contingency planning.

[21] Not [Solomon, 1976], which has a chapter on "Reason and the passions," a section on "The rationality of the emotions," and a chapter on "The logic of the emotions." Not [Minsky, 2006], written by an author who knows something about AI. But work on modeling artificial characters for applications in areas like interactive fiction might be useful; see [Bates, 1994].

Perhaps—to develop the example of preference for white wine a bit further—the only mechanism that is needed to generate conditional desires is the ability to imagine different circumstances, together with the ability to color these circumstances as pleasant (to some degree), and unpleasant (to some degree). But it is unlikely to be this simple, because pleasantness is not monotonic with respect to information: I find the idea of a glass of white wine quite pleasant, but the idea of a glass of white wine with a dead fly in it quite unpleasant. Also, my feelings about some imagined situations can be mixed, with elements that I find pleasant and elements that I find unpleasant. At this point, I might have to invoke a conflict resolution method that has little or nothing to do with the emotions.

This leads to a further point: there is a difference between raw or immediate desires, or *wishes*, and all-things-considered desires, or *wants*. This is because desires can conflict not only with one another, but with beliefs. And, when they conflict with beliefs, desires must be overridden: to do otherwise would be to indulge in wishful thinking.

In [Thomason, 2000], I explored the possibility of using a nonmonotonic logic to formalize this sort of practicalization of desires. The target reasoning consisted of deliberations such as the following. (The deliberator is a hiker who forgot her rain gear.)

1. I think it's going to rain.
2. If it rains, I'll get wet.
3. If I get wet, I'll stay wet unless I give up and go home.
4. I wouldn't like to stay wet.
5. I wouldn't like to give up and go home.

The argument reaches an impasse, and a conflict needs to be addressed to resolve it. There are two possible conclusions here, depending on how the conflict is resolved:

6. On the whole, I'd rather go home.
6′. On the whole, I'd rather go on hiking.

The main purpose of Steps 1–5 is to identify the conflict.

I'm not altogether happy with the theory presented in [Thomason, 2000], but I still believe that the practicalization of desires is an important part of practical reasoning that provides opportunities for using logic to good advantage.

2.5 Intention formation

The product of successful means-end deliberation will be an intention, taking the form of commitment to a plan. But the deliberation would not get started without a goal—and I see no difference between a goal and a provisional (and perhaps very sketchy) intention. Often, even in human agents, these goals come from habits, or from compliantly accepted instructions from other agents.

But sometimes goals arise internally, as outcomes of deliberation. The hiker in Section 2.4 provides an example. If the conclusion of the reasoning is a practicalized desire to turn back and head for home, commitment to the conclusion will produce an intention, which may even become a goal for means-end reasoning. ("How am I to get home?")

This is why practicalization can be an important component of practical reasoning, especially if the reasoner is an autonomous human being.

2.6 What to do now?

Moments will arise in the life of an autonomous agent when there is scope for new activities. These opportunities need to be recognized, and an appropriate task needs to be selected for immediate execution. A busy agent with many goals and a history of planning may have an agenda of tasks ready for such occasions; but even so, an agent may have to stop and think to select a task that is rewarding and appropriate—and this will require reasoning. I do not know if any useful work has been done on this reasoning problem.

2.7 Scheduling, execution and engagement

Some of the examples in Section 1.1 were intended to illustrate the point that there can be deliberation even in the execution of physically demanding, real-time tasks. And there can be such a thing as overplanning, since the plans that an agent makes and then proceeds to perform will need to be adjusted to circumstances.

Also, not all intentions are immediate. Those that are not immediate need to be invoked when the time and occasion are right.

There has been a great deal of useful work on these topics in AI; just one one recent example is [Fritz, 2009].

2.8 Framing a practical problem

Leonard Savage's "Small worlds problem" is replicated in the more qualitative setting of means-end deliberation. A means-end reasoning problem requires (at least) a set of actions, a description of the initial conditions, and a goal. But, even in complex cases, formulations of planning problems don't include every action an agent might perform, or every fact about the current state of the world. Somehow, a goal (like "getting to the airport") has to suggest a method of distinguishing the features of states (or "fluents") and the actions that are relevant and appropriate.

I'm sure that ontologies would be helpful in addressing this problem, but other than this I have very little to say about it at the moment.

References

[Allen *et al.*, 1990] James Allen, James Hendler, and Austin Tate, editors. *Readings in Planning*. Morgan Kaufmann, San Mateo, California, 1990.

[Anscombe, 1958] G.E.M. Anscombe. *Intention*. Blackwell Publishers, Oxford, 1958.

[Asada *et al.*, 1999] Minoru Asada, Hiroaki Kitano, Itsuki Noda, and Manuela Veloso. RoboCup today and tomorrow—what we have learned. *Artificial Intelligence*, 110(2):193–214, 1999.

[Baier and McIlraith, 2008] Jorge A. Baier and Sheila A. McIlraith. Planning with preferences. *The AI Magazine*, 29(4):25–36, 2008.

[Bates, 1994] Joseph Bates. The role of emotion in believable agents. *Communications of the ACM*, 37(7):122–125, 1994.

[Belnap, Jr. *et al.*, 2001] Nuel D. Belnap, Jr., Michael Perloff, and Ming Xu. *Facing the Future: Agents and Choices in Our Indeterminist World*. Oxford University Press, Oxford, 2001.

[Blythe, 1999] Jim Blythe. An overview of planning under uncertainty. In Michael Wooldridge and Manuela Veloso, editors, *Artificial Intelligence Today*, pages 85–110. Springer-Verlag, Berlin, 1999.

[Boutilier *et al.*, 2003] Craig Boutilier, Ronen I. Brafman, Carmel Domschlak, Holger H. Hoos, and David Poole. CP-nets: A tool for representing and reasoning with conditional ceteris paribus preference statements. *Journal of Artificial Intelligence Research*, 21:135–191, 2003.

[Brafman *et al.*, 2006] Ronen I. Brafman, Carmel Domshlak, and Solomon E. Shimony. On graphical modeling of preference and importance. *Journal of Artificial Intelligence Research*, 25:389–424, 2006.

[Bratman *et al.*, 1988] Michael E. Bratman, David Israel, and Martha Pollack. Plans and resource-bounded practical reasoning. *Computational Intelligence*, 4:349–355, 1988.

[Chellas, 1969] Brian Chellas. *The Logical Form of Imperatives.* Perry Lane Press, Stanford, California, 1969.

[Church, 1959] Alonzo Church. *Introduction to Mathematical Logic, Vol. 1.* Princeton University Press, Princeton, 1959.

[Domschlak, 2002] Carmel Domschlak. *Modeling and Reasoning about Inferences with CP-Nets.* Ph.d. dissertation, Ben-Gurion University of the Negev, Be'er Sheva, 2002.

[Doyle, 2004] Jon Doyle. Prospects for preferences. *Computational Intelligence*, 20(2):111–136, 2004.

[Fagin *et al.*, 1995] Ronald Fagin, Joseph Y. Halpern, Yoram Moses, and Moshe Y. Vardi. *Reasoning about Knowledge.* The MIT Press, Cambridge, Massachusetts, 1995.

[Fitting, 2009] Melvin Fitting. Intensional logic. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy.* The Stanford Encyclopedia of Philosophy, Stanford, California, 2009.

[Fox and Clarke, 1991] J. Fox and M. Clarke. Towards a formalization of arguments in decision making. In *Proceedings of the 1991 AAAI Spring Symposium on Argument and Belief*, pages 92–99. AAAI, 1991.

[Fritz, 2009] Christian Fritz. *Monitoring the Generation and Execution of Optimal Plans.* Ph.D. dissertation, University of Toronto, Toronto, 2009.

[Gabbay and Woods, 2005] Dov M. Gabbay and John Woods. The practical turn in logic. In Dov M. Gabbay and Franz Guenthner, editors, *Handbook of Philosophical Logic, Volume XIII*, pages 1–122. Springer-Verlag, Berlin, 2nd edition, 2005.

[Geach, 1963] Peter T. Geach. Imperative inference. *Analysis*, 23, Supplement 1(3):37–42, 1963.

[Georgeff *et al.*, 1999] Michael Georgeff, Barney Pell, Martha Pollack, Miland Tambe, and Michael Wooldridge. The belief-desire-intention model of agency. In Jörg P. Müller, Munidar P. Singh, and Anand S. Rao, editors, *Intelligent Agents V: Agents Theories, Architectures, and Languages*, pages 1–10. Springer-Verlag, Berlin, 1999.

[Goldsmith and Junker, 2008] Judy Goldsmith and Ulrich Junker. Preference handling for artificial intelligence. *The AI Magazine*, 29(4):9–12, 2008.

[Green, 1997] Mitchell Green. The logic of imperatives. In E. Craig, editor, *The Routledge Encyclopedia of Philosophy*, pages 717–21. Routledge, New York, 1997.

[Hansson, 2001] Sven Ove Hansson. Preference logic. In Dov M. Gabbay and Franz Guenthner, editors, *Handbook of Philosophical Logic, Volume IV*, pages 319–394. Kluwer Academic Publishers, Amsterdam, 2nd edition, 2001.

[Harel *et al.*, 2000] David Harel, Dexter Kozen, and Jerzy Tiuryn. *Dynamic Logic.* The MIT Press, Cambridge, Massachusetts, 2000.

[Hofstadter and McKinsey, 1939] Albert Hofstadter and J.C.C. McKinsey. On the logic of imperatives. *Philosophy of Science*, 6:446–457, 1939.

[Jones and Carmo, 2002] Andrew J.I. Jones and José Carmo. Deontic logic and contrary-to-duties. In Dov M. Gabbay and Franz Guenthner, editors, *Handbook of Philosophical Logic, Volume VIII*, pages 265–344. Kluwer Academic Publishers, Amsterdam, 2 edition, 2002.

[Jørgensen, 1937-1938] Jørgen Jørgensen. Imperatives and logic. *Erkenntnis*, 7:288–296, 1937-1938.

[Kahneman and Tversky, 1979] Daniel Kahneman and Amos Tversky. Prospect theory: An analysis of decision under risk. *Econometrica*, 47(2):263–291, 1979.

[Keeney and Raiffa, 1976] Ralph H. Keeney and Howard Raiffa. *Decisions With Multiple Objectives: Preferences and Value Tradeoffs*. John Wiley and Sons, Inc., New York, 1976.

[Kenny, 1966] Anthony J. Kenny. Practical inference. *Analysis*, 26(3):65–75, 1966.

[Lemmon, 1965] Edward John Lemmon. Deontic logic and the logic of imperatives. *Logique et Analyse*, 8:39–71, 1965.

[Lewis, 1979] David K. Lewis. A problem about permission. In Esa Saarinen, Risto Hilpinen, Ilkka Niiniluoto, and Merrill Province Hintikka, editors, *Essays in Honour of Jaakko Hintikka*. D. Reidel Publishing Co., Dordrecht, Holland, 1979.

[Lifschitz *et al.*, 2000] Vladimir Lifschitz, Norman McCain, Emilio Remolina, and Armando Tacchella. Getting to the airport: The oldest planning problem in AI. In Jack Minker, editor, *Logic-Based Artificial Intelligence*, pages 147–165. Kluwer Academic Publishers, Dordrecht, 2000.

[McCarthy and Hayes, 1969] John McCarthy and Patrick J. Hayes. Some philosophical problems from the standpoint of artificial intelligence. In Bernard Meltzer and Donald Michie, editors, *Machine Intelligence 4*, pages 463–502. Edinburgh University Press, Edinburgh, 1969.

[Millgram, 2001] Elijah Millgram. Practical reasoning: The current state of play. In Elijah Millgram, editor, *Varieties of Practical Reasoning*, pages 1–26. The MIT Press, Cambridge, Massachusetts, 2001.

[Minsky, 2006] Marvin Minsky. *The Emotion Machine*. Simon & Schuster, New York, 2006.

[Newell, 1992] Allen Newell. *Unified Theories of Cognition*. Harvard University Press, Cambridge, Massachusetts, 1992.

[NSPE Board of Ethical Review, 2016] NSPE Board of Ethical Review. Case no. 15. https://www.nspe.org/sites/default/files/BER15-1%20APPROVED.pdf, 2016.

[Reiter, 2001] Raymond Reiter. *Knowledge in Action: Logical Foundations for Specifying and Implementing Dynamical Systems*. The MIT Press, Cambridge, Massachusetts, 2001.

[Ros *et al.*, 2009] Raquel Ros, Josep Llus Arcos, Ramon Lopez de Mantaras, and Manuela M. Veloso. A case-based approach for coordinated action selection in robot soccer. *Artificial Intelligence*, 173(9–10):1014–1039, 2009.

[Ross, 1941] Alf Ross. Imperatives and logic. *Theoria*, 7:53–71, 1941. Reprinted with minor changes, in *Philosophy of Science*, Vol. 11 (1944), pp. 30–46.

[Russell and Wefald, 1991] Stuart J. Russell and Eric Wefald. *Do the Right Thing*. The MIT Press, Cambridge, Massachusetts, 1991.

[Sanner and Boutilier, 2009] Scott Sanner and Craig Boutilier. Practical solution techniques for first-order MDPs. *Artificial Intelligence*, 173(5–6):748–788, 2009.

[Savage, 1972] Leonard Savage. *The Foundations of Statistics*. Dover, New York, 2nd edition, 1972.

[Shanahan, 1997a] Murray Shanahan. A logical formalisation of Ernie Davis' egg cracking problem. Unpublished manuscript, Imperial College London, 1997.

[Shanahan, 1997b] Murray Shanahan. *Solving the Frame Problem*. The MIT Press, Cambridge, Massachusetts, 1997.

[Simon and Schaeffer, 1992] Herbert A. Simon and Jonathan Schaeffer. The game of chess. In Robert J. Aumann and Sergiu Hart, editors, *Handbook of Game Theory with Economic Applications, Vol. 1*, pages 1–17. North-Holland, Amsterdam, 1992.

[Simon, 1947] Herbert A. Simon. *Administrative Behavior*. The Macmillan Company, New Yori, 1947.

[Skyrms, 1990] Brian Skyrms. *The Dynamics of Rational Deliberation*. Harvard University Press, Cambridge, Massachusetts, 1990.

[Solomon, 1976] Robert C. Solomon. *The Passions: The Myth and Nature of Human Emotion*. Anchor Press, New York, 1976.

[Thomason, 2000] Richmond H. Thomason. Desires and defaults: A framework for planning with inferred goals. In Anthony G. Cohn, Fausto Giunchiglia, and Bart Selman, editors, *KR2000: Principles of Knowledge Representation and Reasoning*, pages 702–713, San Francisco, 2000. Morgan Kaufmann.

[Thomason, 2002] Richmond H. Thomason. Preferences as conclusions. In Ulrich Junker, editor, *Preferences in AI and CP: Symbolic Approaches*, pages 94–98. AAAI Press, Menlo Park, California, 2002.

[Tversky and Kahneman, 1981] Amos Tversky and Daniel Kahneman. The framing of decisions and the psychology of choice. *Science, New Series*, 211(4481):453–458, 1981.

[Visser and Burkhard, 2007] Ubbo Visser and Hans-Dieter Burkhard. RoboCup: 10 years of achievements and future challenges. *The AI Magazine*, 28(2):115–132, 2007.

[von Wright, 1963] Georg Henrik von Wright. Practical inference. *The Philosophical Review*, 72:159–179, 1963.

[von Wright, 1983] Georg Henrik von Wright. *Practical Reason: Philosophical Papers, Volume 1*. Cornell University Press, Ithaca, 1983.

[Weirich, 2004] Paul Weirich. *Realistic Decision Theory: Rules for Nonideal*

Agents in Nonideal Circumstances. Oxford University Press, Oxford, 2004.

[Williams, 1963] B.A.O. Williams. Imperative inference. *Analysis*, 23, Supplementary 1(3):30–36, 1963.

[Wooldridge, 2000] Michael J. Wooldridge. *Reasoning about Rational Agents*. Cambridge University Press, Cambridge, England, 2000.

Richmond H. Thomason
University of Michigan, USA
Email: rthomaso@umich.edu

7
Deontic Logic and Natural Language

Fabrizio Cariani

Abstract. There has been a recent surge of work on deontic modality within philosophy of language. This work has put the deontic logic tradition in contact with natural language semantics, resulting in significant increase in sophistication on both ends. This chapter surveys the main motivations, achievements, and prospects of this work.

1 Introduction **500**

2 Deontic modality as a linguistic category **500**

3 From conditional obligations to iffy oughts **503**
3.1 Kratzer's theory of conditionals and modals 505
3.2 Chisholm's paradox 509
3.3 The miners paradox 516

4 Puzzles of normality **521**
4.1 Inheritance 522
4.1.1 Disjunction inferences 522
4.1.2 Conjunction inferences 524
4.1.3 Responses and Arguments in Favor of Monotonicity 525
4.2 Agglomeration 527

5 Varieties of deontic strength **530**
5.1 Motivating Data 531
5.2 Accounts of the strength asymmetry. 532

6 The grammar of action **535**
6.1 Logical approaches 536
6.2 Linguistic concerns 537

7 Conclusion **541**

I thank Ana Arregui, Bob Beddor, Janice Dowell, Simon Goldstein, Malte Willer, Aynat Rubinstein, Paolo Santorio, and Malte Willer for conversations, feedback and help with tracking down some papers. I also thank an anonymous reviewer for the Handbook of Deontic Logic for extensive comments, as well as the editors for the invitation to contribute as well as feedback on a previous draft.

Handbook of Deontic Logic and Normative Systems, Volume 2

1 Introduction

The last couple decades have seen a remarkable amount of activity in philosophy of language on the topic of deontic modality. This kind of interaction between philosophy of language and deontic logic is potentially fruitful in both directions. Philosophers of language and linguists leverage their frameworks and techniques to open up new approaches to classic problems in deontic logic. In some cases, problems that appear minor from the perspective of pure logic can become targets of extended analysis in the theory of linguistic meaning. In the opposite direction, the power and sophistication of logical methods can help systematize, constrain and investigate the space of available answers to linguistic questions. While this second direction of cooperation has received less attention, it seems both possible and desirable for it to become more prominent.[1]

In this paper, I survey some of the most striking contributions in this emerging area. This is the structure of this essay: Section 2 clarifies the linguistic scope of the survey — identifying the expressions and concepts that will be in focus; Section 3 surveys work concerning the interaction between deontic modals and conditionals; Section 4 surveys work on a family of puzzles concerning monotonicity properties of deontic operators in natural language; Section 5 takes up linguistic work on the variety of forces for deontic modals–in particular, work on the difference between *ought* and *must*. Section 6 covers work on the relationship between deontic language and the language of agency. The four sections starting with Section 3 are all independent of each other.

Needless to say, all of these topics have direct antecedents in the deontic logic literature. Since my focus is on interactions with research on natural language, I won't chart all the relevant historical references very carefully. Abundant references to these antecedents are available in the other essays in the present handbook — in particular, [Hilpinen and McNamara, 2013].

2 Deontic modality as a linguistic category

Let us start with some rough characterizations. What is a deontic modal? We might follow the opening move in [Portner, 2009] and claim that "modality is the linguistic phenomenon whereby grammar allows

[1] [Holliday and Icard, 2018] make the case for why it's desirable; [Holliday and Icard, 2017] and [Van De Putte, 2018] show a possible shape such work might take.

one to say things about, or on the basis of, situations which need not be real" (p. 1). This seems right but it is not necessarily a great guide when it comes to classifying specific expressions as modals. I propose for the purposes of this essay to complement it with a sufficient condition for modality: classify an expression e as a *modal* (within a given semantic theory T), if e's semantic evaluation rules manipulate a world of evaluation.[2]

Some theorists operate under a definition that is at the same time more precise, stronger, more theoretically loaded, and ultimately more dubious. According to this definition, all modals express concepts of *possibility* or *necessity*. I do not accept this characterization because I think there are modal expressions that are not well understood as either possibility or necessity operators. To give only one example, probability operators are not well understood as possibility or necessity operators [Yalcin, 2010; Lassiter, 2011; Lassiter, 2017]. Additionally, there are analyses of *ought* we will encounter in this essay that deny that *ought* is accurately classified by this scheme. Finally, it is plausible that deontic comparative adjectives are modal in character without being either necessity or possibility operators. Consider for an example the propositional uses of *better* in:

(1) It is better to ride a bike than to drive.

Taken together, these considerations make it preferable to stick with a less committal characterization.

It is an often noted feature of modals that they give rise to a variety of interpretations. Thus, English *may* can convey epistemic possibility (roughly: compatibility with some state of information) or deontic permission (roughly: compatibility with a body of norms). Neither does the list end there: the literature recognizes metaphysical, ability-based, temporal interpretations and many others. That said, not all modals allow every interpretation, and, as we will shortly see, some do not seem polysemous at all.

Deontic interpretations concern modal statuses that obtain in virtue of some norm, or value. Permission and obligation are both prime examples. Following a useful suggestion by [Portner, 2009], we can see deontic interpretations as part of a broader category of *priority* interpretations

[2]The definition ties the class of modals to one's choice of semantic theory. However, we often talk about modals without specifying a theory. For example, English *may* is generally classified as a modal regardless of one's particular theory. I understand (unrelativized) claims that e is a modal as conveying implicit acceptance that e qualifies as modal according to any reasonable choice of semantic theory.

of modals.

> "The idea behind the term "priority" is that such things as rules, desires, and goals all serve to identify some possibility as better than, or as having higher priority than, others." [Portner, 2009, p. 135]

This is a felicitous classification suggestion. Looking at the class of priority modals helps zero-in on some generalizations that are harder to detect when we look at broader classes (e.g., modals generally). And not many new generalizations emerge when we look at more specific classes. For example, there are very few general facts about the deontic *ought* that lack matching facts concerning more goal-driven ones.

There is disagreement about what gives rise to this rich variety of interpretations. A view that is often associated with Kratzer's program is that much of the variety of modal interpretations can be traced down to context-sensitivity.[3] While Kratzer's contextualism is superior to a straightforward ambiguity theory, there is a growing intellectual demand for an account of the diversity of interpretations of modals that derives them on the basis of more systematic consideration.

An alternative is to derive the different modal interpretations on the basis of structural facts. Hacquard [2006, 2010] pioneered an approach to do so within a framework that blends elements of Kratzer's approach with event-semantics and emphasizes the syntactic differences between modal sentences carrying epistemic interpretations and ones that carry non-epistemic interpretations.

Many of the questions that are discussed in this survey are, to a degree, independent of one's views on this matter. But no story about deontic modality in natural language is complete without some account of it.

As I anticipated, it is also important to recognize that not all modals are polysemous. Deontic concepts can be expressed by modal verbs like *require*, *obligate*, *permit*, as well as their nominalizations. These verbs do not generally give rise to the broad variety of interpretations that is associated with, say, *must*. In this connection, [Hacquard, 2013] distinguishes between *grammatical* modals and *lexical* modals. The hallmarks of grammatical modals (like *must*, *ought*, *may*, *can*, etc.) are that they are closed class expressions and are polysemous. ([Szabó, 2015] suggests that the category of closed class expressions is a useful proxy in natural

[3] [Kratzer, 1977; Kratzer, 1981; Kratzer, 1991b; Kratzer, 2012]. For rich developments of the contextualist view of modal flavors see also [Dowell, 2011; Dowell, 2013; Bronfman and Dowell, 2018; Bronfman and Dowell, 2016].

language for what logicians call "logical constants".) By contrast *lexical* modals (*likely*, *obligatory*, *permitted*) are open class expressions and are typically not polysemous.[4]

The research surveyed here focuses heavily on grammatical modals, and so on the deontic interpretations of *must*, *ought*, and *may*. This is no doubt in part for contingent reasons and in part because the closed class expressions appear closer to the architectural features of a language. Despite that, grammatical priority modals are not insulated. They are embedded in inferential networks that relate them to lexical modals as well. So, our discussion will occasionally touch on lexical modality.

3 From conditional obligations to iffy oughts

A historically important tradition in deontic logic focuses on the concept of *conditional obligation*—what one is obliged to do given that some condition holds. Philosophers of language and linguists have revised and remixed some of the main arguments in this tradition. Their distinctive concerns have led to substantial progress, as well as to the opening of some new avenues of inquiry.

Let us take off our exposition from a famous passage in David Lewis's "Semantic Analyses for Dyadic Deontic Logic" [Lewis, 1974]. Keep in mind that, at this point in time, Lewis is already summarizing a wealth of prior work on conditional obligation.

> "It ought not to be that you are robbed. A fortiori, it ought not to be that you are robbed and then helped. But you ought to be helped, given that you have been robbed. The robbing excludes the best possibilities that might otherwise have been actualized, and the helping is needed in order to actualize the best of those that remain. Among the possible worlds marred by the robbing, the best of a bad lot are some of those where the robbing is followed by helping." [Lewis, 1974, p.1]

Lewis took these considerations to motivate an analysis of dyadic conditional obligation. As will no doubt be familiar, standard deontic logic is built on the idea that obligation operators are unary necessity operators. Following an extended analogy with conditional probability, we

[4]This distinction has heuristic value but it is unlikely to be perfectly clear cut, since there are evidently lexical modals that do appear to be polysemous. For example, *warranted* and *justified* seem to have both epistemic and deontic interpretation, while *compel* seem to have both causal and deontic interpretations.

might step away from that paradigm and instead introduce *binary* obligation operators. For instance, the operator $\bigcirc(\mathsf{Q} \mid \mathsf{P})$ could be used so as to mean that Q is obligatory given that we restrict focus to the possibilities satisfying P. Lewis's example could be modeled as follows: $\bigcirc(\textit{you are helped} \mid \textit{you are robbed})$.

Studying the logic of such dyadic operators is an entirely unobjectionable enterprise, especially if it is divorced from considerations of natural language. However, many authors have pointed out that, from the point of view of our understanding of natural language, it is a mistake to think of *iffy oughts* like (2) as having the form $\bigcirc(\textit{nap} \mid \textit{tired})$.[5]

(2) If Iris is tired, she should take a nap.

One reason for resisting the representation in terms of conditional obligation operators is that it misses out on the generality of conditional modality. For one thing, we may want to analyze sentences with multiple operators in their consequent, such as:

(3) If Iris is tired, she will try to stay awake but she should take a nap.

For another, each deontic modal expression would need its own dyadic operator. After all, we can have conditional permissions as in (4) and other kinds of conditional deontic claims as in (5).

(4) If you have the permit, you may fix your sidewalk.

(5) If you bought this guitar, you must buy this amplifier.

Setting generality aside, from the point of view of theory-design, the idea of dyadic obligation operators seems insufficiently modular. If the point of the formal semantic system is to contribute to a compositional theory of meaning for a language like English, it should separate the contributions of *if* and *ought* (*viz.*, *if*+*may* and *if*+*must*, depending on the case). As Thomason put the point:

> "A proper theory of conditional obligation [...] will be the product of two separate components: a theory of the conditional and a theory of obligation." [Thomason, 1981b, p.165]

Here is Bonevac riffing on Thomason's theme:

> "At the very least, a theorist using a conditional obligation operator owes us an explanation of how the semantics of

[5] I borrow the 'iffy ought' terminology from [Willer, 2012].

> the operator depends on the semantics for obligation and the conditional simpliciter. Sentences expressing conditional obligations are intelligible to anyone understanding *should* (or *ought to*) and *if*. The combination of these words is no idiom. The meanings of such sentences, therefore, should be explicable in terms of the meanings of *if* and *should* construed independently." [Bonevac, 1998, p.37]

Thomason and Bonevac's comments seem entirely right to me. Anyone who cares about a formal theory of meaning should heed their advice and reject conditional obligation theories. The next question then is what *do* we learn by separating out the contributions of *if* and *should*?

3.1 Kratzer's theory of conditionals and modals

Proponents of conditional obligation operators lost a battle on the interpretation of deontic conditionals like (2), but they ended up scoring an unexpected win in a war they might not even have known they were fighting.

The semantic framework for modality that has come to be viewed as paradigmatic, Kratzer's,[6] incorporates some of the key ideas of the dyadic analysis—while also heeding Thomason's demand for a compositional analysis of sentences like (2). In light of its prominence, and of the fact that much of the literature reacts to this paradigm, I will present a simplified version of Kratzer's theory.[7] I will refer to this as the *baseline* theory.

According to the baseline theory, every modal (and crucially, the grammatical modals *ought*, *must*, *may*) takes two propositional arguments, a restrictor and a scope.[8] The scope proposition is the proposition whose modal status we are interested in. The restrictor proposition delimits the set of worlds that are to count as relevant. Thus in (6), the restrictor is the proposition that you have a permit, the scope is the proposition that you fix your sidewalk.

(6) *may*[you have a permit][you fix your sidewalk]

The aggregate interpretation is something like what we'd express by saying *if you have a permit, you may fix your sidewalk*. Crucially, even

[6] [Kratzer, 1977; Kratzer, 1981; Kratzer, 1991a; Kratzer, 1991b; Kratzer, 2012].

[7] It should be noted that Kratzer has explored, advocated and developed a variety of other frameworks and revisions of her theory.

[8] Modals might, in fact, take more than two arguments depending on the details of the semantic theory. I ignore this complication.

the overtly unconditional *you may fix your sidewalk* gets assigned a logical form like:

(7) *may*[RESTRICTION][you fix your sidewalk]

In this representation, RESTRICTION is a place-holder for a restrictor proposition. Restrictor propositions are provided by context, and usually are a bit more generic than the kind of antecedent in (6). Plausibly (7) is interpreted relative to a restriction to the worlds in which the city laws and the relevant circumstances are about the same as they are in the base world. The crucial point is that, according to the baseline analysis, restrictor propositions are *always* needed to interpret modal claims. So far so good: the dyadic obligation theorist has been nodding all along.

Kratzer adds a few important theses. To start, the theory is to apply to *all* modals, and not limited to some specific interpretation (Kratzer uses "flavors" to refer to what I have described as "interpretations" of modals—i.e., epistemic, deontic, etc.). Furthermore, as I noted in the previous section, these different interpretations arise as a result of different settings for various contextual parameters.

These parameters include the parameters that fix the contextual restrictor proposition. Critically, however, the restrictor proposition may also be affected by explicit elements of the sentence, and perhaps even elements of the discourse at large. The principal (but not the *only*) device of restriction is the conditional *if*. Let us see how all of these ideas play out in the formalism.

I said that the restrictor proposition is partly determined by context. In the baseline theory, context provides values to two parameters:

- a *modal base* f (mapping the set of all worlds Ω to $\mathcal{P}(\Omega)$)
- an *ordering source* g (mapping Ω to preorders $\Omega \times \Omega$)

If fed a world of evaluation w these functions output a set of worlds M (the *modal background relative to* w) and a pre-order of worlds $\prec$.[9]

[9]Kratzer's official theory is formulated in the framework of premise semantics. This means that the types of these contextual parameters are slightly different from what I have suggested. In particular, modal bases are actually functions from worlds to sets of propositions ($\Omega \mapsto \mathcal{P}(\mathcal{P}(\Omega))$) which then determine a set by intersection. Similarly, her ordering sources are also functions from worlds to sets of propositions, which then determine a pre-order. The premise semantics formulation of the theory has some explanatory advantages that will come in handy in Section 5. But for now it will be quicker to set it aside.

The concepts of modal base and ordering source help provide lexical entries for modals. These entries are easiest to state under restrictions that entail the *limit assumption* (see below for the content of this assumption). For example, assume that there are finitely many possible worlds. Then we can give the following analyses for the modals *must*, *might* and *ought*. (We momentarily assume that *ought* and *must* have the same meaning, but we will question this assumption in Section 5.)

$$\textsc{best}(f,g,w) = \{u \in f(w) \mid \neg\exists v, v \succ_{g,w} u\}$$

(8) $$[\![ought]\!]^w = [\![must]\!]^w = \lambda\mathbf{P}.\lambda f.\lambda g.\forall v \in \textsc{best}(f,g,w), v \in \mathbf{P}$$

(9) $$[\![might]\!]^w = \lambda\mathbf{P}.\lambda f.\lambda g.\exists v \in \textsc{best}(f,g,w) \;\&\; v \in \mathbf{P}$$

To illustrate these clauses, consider evaluating *you must run* at world w. This will be true (at w) if every world in $\textsc{best}(f,g,w)$ is a world in which you run. The worlds in $\textsc{best}(f,g,w)$ are exactly those worlds in $f(w)$ such that no world is better than them (in the sense of $\prec_{g,w}$). If $f(w)$ contains a world in which you walk that is better than every world in which you run, our target sentence will be false. It will also be false if for every run-world there is a better walk-world (whether or not there is a walk-world that beats all the run-worlds). Otherwise, it'll be true.

The entries above presuppose the infamous *limit assumption*. Part of the content of the limit assumption is the claim that, for every choice of $f(w)$, there are maximal points in the partial order $\succeq$.[10] However, it seems plausible, given the deontic interpretation of $\succeq$, that this property could sometimes fail to hold of a deontic ordering. For example, there could be setups that contain infinite chains of worlds with monotonically increasing value. To address that possibility, Kratzer proposes a somewhat more complicated quantificational condition.

$$\textsc{forceable}(\mathbf{P},f,g,w) = \{u \in f(w) \mid \exists v \in f(w), v \succeq_{g,w} u \;\&\; \forall z \in f(w), \textit{if } z \succeq_{g,w} v, z \in \mathbf{P}\}$$

(10) $$[\![ought]\!]^w = [\![must]\!]^w = \lambda\mathbf{P}.\lambda f.\lambda g.\forall v \in f(w), \textsc{forceable}(\mathbf{P},f,g,v)$$

In words, $\mathbf{P}$ is forceable from u's perspective if there is a relevant world v that is at least as good as u such that all of the relevant worlds that are at least as good as v make $\mathbf{P}$ true. And *must*(P)/*ought*(P) is true if the prejacent proposition $\mathbf{P}$ is forceable from the perspective of any

[10] I take it that the content of the limit assumption is not exhausted by this condition. See [Kaufmann, 2017] for an extensive study, detailing and resolving much outstanding confusion about how the limit assumption ought to be formulated.

relevant world. [11]

The limit assumption is an excellent example of a useful idealization. It is almost certainly false, given the intended interpretation of the ordering. But, in nearly every application, working under more realistic assumptions increases the cognitive load for modest benefit. In general, it is practical to default to working under conditions that support the limit assumption.

The baseline theory of modals is complemented, and completed, by Kratzer's theory of conditionals. As noted, in Kratzer's semantics, *if* does not denote a binary connective. It serves instead to further restrict the modal base of a modal in its scope. It is not straightforward to give a fully compositional implementation of this idea,[12], so we will keep it at a relatively intuitive level. The *effect* of this idea is that sentences that look like $(\mathit{if}\ \mathsf{P})(\bigcirc\mathsf{Q})$ are evaluated by:

(i) interpreting the restrictor proposition (i.e. whatever proposition $\mathbf{P}$ is expressed by P in context);

(ii) restricting the modal base f for $\bigcirc$ with $\mathbf{P}$ (i.e. creating a new function $f + \mathbf{P} = \lambda w.f(w) \cap \mathbf{P}$);

(iii) evaluating the modal claim, $\bigcirc Q$ relative to this shifted modal base $f + \mathbf{P}$ and whatever ordering source g was provided by the initial context.

Of course, some *if*-sentences do not have overt modals. In these cases, Kratzer's hypothesis is that *if* restricts a *covert* modal. This covert modal defaults to an epistemic necessity interpretation. In other words (11) is actually interpreted as (12) with the restrictor proposition narrowing down the modal base of $\Box$.[13]

[11]Let us consider why (10) and (8) are equivalent in the special case in which there are finitely many worlds. (Recall that this is not the only hypothesis compatible with the limit assumption, but it will be illustrative regardless of that.) Suppose *must* P meets the truth-condition in (8); then all the worlds such that nothing is better than them are P-worlds. But so, if we consider an arbitrary relevant world v we should be able to find whichever world that is identical to or better than v that is maximal and let that world be the witness to the existential quantifier in the definition of FORCEABLE. In the opposite direction, suppose that the proposition expressed by P is FORCEABLE relative to f and g and any relevant world. Consider an arbitrary world v belonging to $\textsc{best}(f, g, w)$, then v is in $f(w)$ so P is forceable with respect to v. However, since v is a terminal world with respect to $\succ_{g,w}$, all that forceability entails is that P must be true at v as demanded by the analysis.

[12]See [von Fintel, 1994, ch.3] for an example of the sort of work that is involved.

[13]This opens up the possibility that there might also be cases in which a covert

(11) If she called, she lost.

(12) If she called, □ (she lost).

With the baseline view developed let us see how it applies to some classic problems from deontic logic.

3.2 Chisholm's paradox

The baseline theory makes a distinctive prediction about the classical paradoxes of conditional obligation. I illustrate the prediction by considering Chisholm's paradox [Chisholm, 1963]. The theory makes similar predictions in the case of the Good Samaritan paradox and Forrester's Gentle Murder paradox [Forrester, 1984]. I will then compare the baseline with some alternatives.

Here is a slightly touched-up version of Chisholm's familiar vignette:

> Your elderly neighbors asked for your assistance to prepare their taxes. Because you are their only acquaintance, *you ought to help them*. But, they are easily frightened by potential intruders. So *it ought to be that, if you go, you tell them in advance*. However, telling them that you are going will be bad if you are *not* going. So, *if you don't go, you ought not to tell them*. As it happens, your favorite show is on TV and *you don't go*.

The four italicized sentences generate the puzzle. To get a little closer, we can do a preliminary formalization of these:

(i) $\bigcirc(go)$

(ii) $\bigcirc(if\ go,\ tell)$

(iii) $if \neg\ go, \bigcirc(\neg\ tell)$

(iv) $\neg go$

This formalization is not meant to be definitive. Depending on our background commitments, we might want to revise aspects it. For example, defenders of the baseline theory might want to revise (ii) in ways that will become clear shortly.

modal is posited in addition to an overt one. This possibility is leveraged in [Geurts, 2004] and discussed further with application to conditional deontics in [Cariani *et al.*, 2013].

Whatever the final formalization, critical elements in the paradox are already apparent at this preliminary level. Chisholm's key assumption is that $\bigcirc$ scopes outside the conditional in (*ii*) but inside the conditional in (*iii*).

Chisholm noted that these sentences sound collectively consistent. The puzzle is that they fail to be consistent under standard assumptions of deontic logic, combined with two principles governing the interaction of conditional and obligation operators. The two principles are "factual" and "deontic" detachment. It is indeed common (though perhaps not entirely correct) to frame Chisholm's puzzle as pitching these against each other [Loewer and Belzer, 1983].

FACTUAL DETACHMENT (FD). $\mathsf{P}, (\mathit{if}\ \mathsf{P},\ \bigcirc\mathsf{Q}) \vdash \bigcirc(\mathsf{Q})$

(WIDE) DEONTIC DETACHMENT (DDW). $\bigcirc(\mathsf{P}), \bigcirc(\mathit{if}\ \mathsf{P}, \mathsf{Q}) \vdash \bigcirc(\mathsf{Q})$

Though I formulate deontic detachment with $\bigcirc$ taking with scope over a conditional, it is also worth considering a version of deontic detachment in which the deontic modal appears in the consequent of the conditional.

NARROW DEONTIC DETACHMENT (DDN). $\bigcirc(\mathsf{P}), (\mathit{if}\ \mathsf{P}, \bigcirc\mathsf{Q}) \vdash \bigcirc(\mathsf{Q})$

After all, a defender of the restrictor analysis might formulate the second premise of Chisholm paradox as: *if go*, $\bigcirc(\mathit{tell})$

Chisholm's formulation of the paradox indicates that he intends to construe premise (ii) with the conditional scoping under the deontic modal. That would give relevance to DDW as far as the argument goes. However, as [Saint Croix and Thomason, 2014] forcefully note, there are very few constructions in English that are appropriately represented with a deontic modal scoping over a conditional. In the following, I use DD for points that go through on either construal of this principle.

Under the assumption that sentences of the form $(\mathit{if}\ \mathsf{P},\ \bigcirc\mathsf{Q})$ are conditionals with $\bigcirc$ in the consequent, FD just is *modus ponens*. As we saw, this assumption is strictly speaking false on the baseline view, since on the baseline view conditionals restrict modals, as opposed to *connecting* pairs of propositions. To minimize our committments, it is be preferable to stick to calling it "factual detachment".

Here is a sketch of the proof of inconsistency, taking the four sentences as premises:

(v) $\bigcirc(\mathit{tell})$ (i),(ii) DD

(vi) $\bigcirc(\neg\mathit{tell})$ (iii),(iv) FD

(vii) $\bigcirc(tell \ \& \ \neg tell)$ (v),(vi) AGGLOMERATION

(viii) $\bot$ (vii), AXIOM D

There are many familiar options to address the paradox. Giving up deontic detachment, giving up factual detachment for *iffy* oughts; rejecting agglomeration;[14] giving up some structural rules; arguing that the premises need to be formalized differently. The question that matters for present purposes is which of these answers best fits with the constraints and commitments coming from various theories within natural language semantics.

For example, I mentioned above that the baseline theory makes a distinctive prediction. In particular, the theory allows all four sentences to be true at once by blocking FACTUAL DETACHMENT. To see this, model the case by assuming that there are four worlds in the modal base $w_{GT}, w_{G\overline{T}}, w_{\overline{G}T}, w_{\overline{GT}}$.[15] Suppose that, as is plausible, these are ranked as follows in the relevant ordering source:

$$w_{GT} > w_{G\overline{T}} > w_{\overline{GT}} > w_{\overline{G}T}$$

This parameter setting has the following effects. The unconditional claims $\bigcirc(go)$ and $\bigcirc(tell)$ are true because w_{GT} is the best world. For the same reason, $\bigcirc(\neg tell)$ is false. If the second premise is interpreted along the lines of DDN, it is easy to see that the conditional antecedent restricts the domain to the two top-most worlds. Among these, the best is the one in which you tell.[16] For the third premise, processing the conditional antecedent *if you don't go* restricts the modal background to

[14] If we assume the duality of obligation and permission, the agglomeration step could be avoided in the proof. We might reason from (vi) to $P(\neg tell)$ via axiom D which would directly contradict $\neg P \neg tell$, which is equivalent to (v). In general, the strategies I will mention below that block the agglomeration step, must also reject duality.

[15] The indices are designed to convey which of *go* and *tell* are true at each world.

[16] Things are a bit more complicated if we formalize the premise as DDW. If we have to stick to the idea of wide-scoping $\bigcirc$, the baseline predicts the following logical form for premise (ii):

$$\bigcirc(\text{RESTRICTION})(if\ go, \Box tell))$$

In this, RESTRICTION picks out the contextual restriction (which in this case we assume to be tautological, so that no worlds are ruled out). Whether premise (ii) is true, will depend on what kind of modal $\Box$ is. If its domain at each world is simply $\{w_{GT}, w_{G\overline{T}}, w_{\overline{G}T}, w_{\overline{GT}}\}$, then the sentence will be false. It could be made true by forcing the modal $\Box$ to have stricter domains. For example, the modal base and ordering source of $\Box$ could be set up so that $if\, go, \Box tell$ is equivalent to the material conditional $go \supset tell$. This could be accomplished for instance by requiring that for

$\{w_{\overline{GT}}, w_{\overline{G}T}\}$ world. The set of best worlds in this restricted background is the singleton $\{w_{\overline{GT}}\}$. So, under the restriction to $\neg go$, $\bigcirc(\neg tell)$ is true. For the fourth premise, just assume that the actual world is either of the $\overline{G}$-worlds.

The upshot: the baseline validates DD, AGGLOMERATION, AXIOM D but invalidates FD.

One worry about this type of approach, articulated in [Arregui, 2010], is that part what's puzzling about Chisholm's case can be replicated without at all involving FD. Indeed, we have a lingering, and hiterto unaddressed, intuition that the relevant obligations one is under might change with time, and specifically might depend on how some things turn out in the actual world. Once it is settled that you won't go to help your neighbor, it is perhaps no longer the case that you ought to go. The point is best illustrated with some nearby cases. Following Arregui, consider the argument from (13)-(14) to (15):

(13) It ought to be the case that Nina does her job well.

(14) It ought to be the case that if Nina does her job well, she gets a promotion.

(15) It ought to be that Nina gets a promotion.

Surely whether Nina gets a promotion does not just depend on the purely normative premises (13) and (14). Whether she gets a promotion should also depend on how things are in the actual world—for example, on whether she does her job well.[17]

The theoretical proposal of [Arregui, 2010] involves a solution to Chisholm's paradox that involves giving up *both* DD and AGGLOMERATION. I will present Arregui's semantics in some detail, for it is extraordinarily interesting and somewhat neglected in the philosophical literature.

Arregui presents her preferred semantics for $\bigcirc$ within the framework of situation semantics.[18] Start with the idea that possible worlds have parts. Call these parts *situations*. Assume that situations are structured by a relative parthood relation which is reflexive, antisymmetric and

each world w, w is the unique highest ranked modal in the ordering source of $\Box$ (even if it isn't the highest ranked world in the ordering source of $\bigcirc$). Simulating the material conditional within the baseline view approximates the interpretation that Chisholm had in mind for sentence (*ii*) and also makes it true relative to our chosen parameter settings.

[17] See also [Lassiter, 2017, §8.6] for a discussion that echoes these points.

[18] I don't think this is *essential* but it certainly is heuristically helpful.

transitive. Possible worlds are maximal situations with respect to the parthood relation. Say that a situation s *deontically requires* P if all the deontically best ways of extending s into a possible world make P true. My promising to help you requires my helping you: this is because in every ideal world that extends the situation of the promise I help you.

Now, suppose that you want to evaluate $\bigcirc(\mathsf{P})$ at world w. Consider any situation s that is (i) part of w (ii) compatible with P and (iii) not enough to guarantee P on its own. Say that these are the *P-relevant* situations in w.[19] With these concepts in hand we can state Arregui's semantic theory:

(16) $\bigcirc(\mathsf{P})$ is true at w if every s that is P-relevant in w is part of a situation s' that requires P.

Note the double layer of quantification: every P-relevant situation has to be extendable to a situation s' that requires P.[20]

We can illustrate these ideas by considering their application to (15) — *It ought to be that Nina gets a promotion.* This is predicted to be false in a world in which Nina does not do her job well and true in a world in which she does. To see why it's false in worlds in which Nina does not do her job well, let s be a situation that includes her accepting the job and every instance of Nina doing the job poorly. Intuitively, there is no way of extending this situation into one that requires that she be promoted. To see why it's true in worlds in which Nina does does her job well, start with any situation s, and extend s to a situation s' that includes the situation of Nina doing her job well. It is intuitive that such a situation s' deontically requires her getting the promotion— and plausibly this requirement will be formalized once we have a correct theory of premise (14).

For conditional oughts, Arregui proposes an account that works out to the following truth-conditions. To evaluate $(\mathit{if}\,\mathsf{P})(\bigcirc\,\mathsf{Q})$ at w, consider all the P-relevant situations s in w; the claim will be true if every such s can be extended to an s' that in turn can be extended to a minimal P-satisfying s'' such that s'' requires P. For (14), imagine extending Nina's

[19] Arregui's situations are world-bound, in the sense that they belong to at most one world, though they may have counterparts in other worlds.

[20] A moment's reflection should clarify why the concept of P-relevance is needed: situations incompatible with P are not extendable into possible worlds that verify P, whether ideal or not; and the situations that guarantee P are trivially extendable.That said, omitting condition (iii) would probably not affect the truth conditions. Since (16) quantifies universally over P-relevant situations, it'll do no harm to include some that are trivially part of situations that require P.

accepting the offer to a situation s'' in which she does her job well, then s'' requires Nina's getting a promotion, making the conditional true.

It is an important feature of this theory (though not one that is explicitly discussed in Arregui's paper) that it invalidates both DD and AGGLOMERATION of $\bigcirc$ over conjunction. That is, the proof in (*v*)-(*viii*) contains two invalid steps (given the semantics), namely (*vi*) and (*vii*). We might find it surprising that the solution is not minimal. If Chisholm's paradox is to drive us away from some ideal of classicality, some might prefer to be driven as little as possible.

However, it turns out that the non-minimality of Arregui's solution to Chisholm's paradox is a crucial feature, and not a bug. This is illustrated by another recent observation in the literature: no completely satisfactory solution to Chisholm's paradox can be limited to abandoning DEONTIC DETACHMENT. This was initially pointed out by [Saint Croix and Thomason, 2014], and later echoed by [Fine, unpublished]. These authors show that there are versions of Chisholm's paradox that do not involve DD at all. Consider this:

> You ought to go help your elderly neighbors. But if you don't help them, you ought to apologize to them. However, your favorite show is on TV and so you don't go.

From the second and third premise, derive that you ought to apologize via FD. From this, via AGGLOMERATION, derive that you ought to go *and* apologize. While that is not itself contradictory, it is a rather odd consequence that involves only FD and AGGLOMERATION. In sum, giving up DD cannot be the entire story.

The accounts of Chisholm's paradox we have reviewed until now require giving up at least one of FD or DD. But another angle that has received substantial exploration in the philosophy of language literature concerns the prospects for an account of Chisholm's paradox that gives up *neither* of these. Views with this feature have been explored independently and with very different tools by [Willer, 2014] and [Saint Croix and Thomason, 2014], in philosophy of language. These works parallel a related tradition that explores the same conceptual space within deontic logic.[21]

The essence of both these approaches is that the fourth premise (*you don't go*) changes the context in a way that undermines the first (*you ought to go*). These views accept that $\bigcirc(\neg tell)$ does (in some sense)

[21] [Prakken and Sergot, 1996; Prakken and Sergot, 1997; Jones and Pörn, 1985; Carmo and Jones, 2002; Gabbay, 2012].

follow from (*iii*) and (*iv*) and $\bigcirc$(*tell*) does (in some sense) follow (*v*) and (*vi*). Modeling the context-shift allows us to account for why all four sentences seem acceptable while denying that they all are acceptable in one and the same context. The two models differ in how they represent the context shift and in what relations they bring to bear to the account of inference patterns.

Willer models this insight in the context of update semantics (extending work by [Veltman, 1996] to the deontic case): a very rough way of putting his view is that he thinks that (a) FD and DD are both valid; (b) however valid patterns of inference are defeasible. In the specific case, the application of DD at step (*v*) is disallowed because the conditional *ought* in premise (*ii*) is defeated by the information in premise (*iv*), i.e. that you don't go.

By contrast, Saint-Croix and Thomason develop their account in a more standard contextualist static semantics. Strictly speaking, in fact, Saint-Croix and Thomason do not *validate* DD. Instead, on their proposal it turn out to be "pragmatically valid" roughly in the sense in which [Stalnaker, 1975] talks about "reasonable inference". One may reasonably expect that natural language inferences are not accounted for purely in terms of an unaided concept of logical validity. Instead, we supplement our arsenal with a pragmatic notion that tracks something like preservation of acceptability. What Saint-Croix and Thomason suggest is that DD (and in particular DDN) meets this criterion.

To wrap up this discussion, I want to draw attention to some general morals about the significance of this debate, rather than probe the individual views. The important point, which seems undeniable to me, is that linguistically oriented investigation into Chisholm's paradox has substantially deepened our understanding of the paradox. We have a better sense of the constraints on a solution to it and of the tradeoffs involve in various approaches. In particular, the collective moral of the papers I have just surveyed is that a complete solution to Chisholm's paradox must at least:

(1) explain why some *ought*-claims seem to depend on what is actual,

(2) explain why contrary-to-duty oughts sometimes cannot be conjoined felicitously with unconditional ones, and

(3) be faithful to the particular way in which *ought*-sentences are context-dependent.

In addition, this investigation has improved our understanding of Kratzer's semantics. It is only as part of work like this that researchers

have started to emphasize the fact that Kratzer semantics invalidates *modus ponens* — or at least, what on the surface looks like *modus ponens*. Applying theories of the linguistic meaning of modals to well-known problems in logic turns out to be a remarkable way of gaining insight about both.

3.3 The miners paradox

Philosophers of language have also infused fresh blood into the discussion of the *miners paradox*. While the history of the miners paradox places it more squarely within the tradition of metaethics rather than that of deontic logic,[22] the paradox does have implications for deontic logic, and particularly so in the revamped version that has become the center of so much recent research.

This resurgence started with an influential article by [Kolodny and MacFarlane, 2010]. They represent the miners scenario as follows:

> "Ten miners are trapped either in shaft A or in shaft B, but we do not know which. Flood waters threaten to flood the shafts. We have enough sandbags to block one shaft, but not both. If we block one shaft, all the water will go into the other shaft, killing any miners inside it. If we block neither shaft, both shafts will fill halfway with water, and just one miner, the lowest in the shaft, will be killed."

The paradox is this: intuitively, all of the following sentences seem acceptable in the given scenario.

(17) Either the miners are in shaft A or in shaft B.

(18) If the miners are in shaft A, we should block shaft A.

(19) If the miners are in shaft B, we should block shaft B.

(20) We should block neither shaft.

The joint acceptability of these sentences suggests that they ought to be consistent, given a semantic theory for *ifs* and *oughts*.[23] However,

[22] Previous work that is of significance here is [Regan, 1980; Parfit, unpublished; Jackson, 1991; Goble, 1996].

[23] The data-point is sometimes presented as the claim that all of (17)-(20) are *true*. Since there are views on which *ought*-sentences and conditional sentences do not strictly speaking have truth-conditions (in the sense that they do not divide the worlds in which they are true from the worlds in which they are false), I prefer the less commital terminology of consistency and joint-acceptability. These are concepts

they turn out to be inconsistent under basic assumptions. In particular, they are inconsistent if *if* and *or* are both interpreted along the lines of classical logic. And, perhaps more importantly, they are inconsistent if the sentences are modeled in the context of the Kratzerian baseline view (even if its account of *if* is non-classical in the ways I have outlined above).

Kolodny and MacFarlane put abundant effort into rejecting the tempting, but mistaken, idea that the Miners Paradox is easily resolved by distinguishing between subjective and objective oughts. They are right that it is not.

More positively, they offer a two-part diagnosis for what goes on in the Miners Paradox. The first part of the diagnosis is that the paradox forces us to invalidate *modus ponens*. The second is that the paradox requires us to adopt a *seriously information sensitive* analysis of *ought* and *should*. (I will explain shortly what this means.)

The *modus ponens* diagnosis is widely viewed as off the mark for two reasons. First, it suggests that according to received views *modus ponens* is valid. This is not so on the Kratzerian baseline view (which has as good a claim as any other view to the title of "received view"). As we saw in the discussion of Chisholm's paradox, the baseline invalidates *modus ponens*. However it still makes the foursome inconsistent (as was independently pointed out in [Charlow, 2013] and [Cariani *et al.*, 2013]).

More importantly, invalidating *modus ponens* isn't essential to the solution of the Miners Paradox. There are sensible theoretical packages that hold on to *modus ponens* and instead give up some of the inference patterns that depend on subarguments. For example, one can give up the combination of *disjunctive syllogism, modus tollens* and *reductio ad absurdum*.[24] The upshot of these two considerations is that the validity of *modus ponens* doesn't really matter all that much to an understanding of the Miners Paradox.

What does matter is the point about information sensitivity. Recall how I described the baseline algorithm for evaluating claims of the form *if* $\mathsf{P}, \bigcirc\mathsf{Q}$. Very roughly:

(i) Ask context for the modal background $f(w)$ of $\bigcirc$ and for some kind of preorder $\succeq_w$ of worlds by their relative priority.

that even a non-truth-conditional theorist would need to have a theory of (though perhaps a non-classical theory).

[24] Various facets of this point are highlighted in [Willer, 2012; Yalcin, 2012; Bledin, 2015; Bledin, 2020]. If the trio of principles above seems to be a jerry-rigged collection, note that they are exactly those principles that are typically associated with local assumptions and subproofs in natural deduction systems.

(ii) Use these elements to compose a pre-domain d consisting of the best (according to $\succeq$) worlds in M.

(iii) Evaluate P and compose the final domain D by restricting d with $\mathbf{P}$ (i.e. $D = d \cap \mathbf{P}$).

(iv) Check that all worlds in D verify Q.

The Miners scenario illustrates a defect in this algorithm. The problem is that the ordering of worlds should be sensitive to the shifts that are introduced by conditional supposition. It may be unconditionally best to block neither shaft. But if the miners are in shaft A, it will be best to block A; and if they are in shaft B it will be best to block B. This is not allowed by the baseline algorithm: if w is the best world in some initial information state, then it remains best under any supposition P that is true at w [Charlow, 2013; Cariani *et al.*, 2013].

What makes the semantics information sensitive is that it satisfies a strengthening of the general principle that what one ought to do in an initial information state i, need not be what one ought to do under information states stronger than i. The condition is most simply put in terms of domains of quantification (though the idea of an information sensitive semantics is not limited to quantificational accounts of *ought*). Any quantificational semantics for *ought* should provide some kind of function, call it `best` that given an information state i and whatever other general parameters π outputs a domain for *ought*-sentences. The core claim of a quantificational information sensitive semantics is this:

INFORMATION SENSITIVITY it is not the case that if i^+ refines i and i^+ is consistent with $\texttt{best}(i, \pi)$, then $\texttt{best}(i^+, \pi) \subseteq \texttt{best}(i, \pi)$

The bug in the baseline domain determination algorithm is that it only allows the set of ideal worlds to change when the new information rules out *all* of the formerly best worlds.

As for satisfying INFORMATION SENSITIVITY, there are now a plethora of proposals, all of which are distinguished by their own set of advantages and disadvantages.[25] One question in this domain that should perhaps receive more attention is how to reconcile the idea that subjective oughts are sensitive to information with the rather common idea within semantics that deontic modals take *circumstantial* modal

[25]In addition to [Kolodny and MacFarlane, 2010], see also [Cariani *et al.*, 2013; Cariani, 2016b; Carr, 2015; Charlow, 2013; Charlow, 2016; MacFarlane, 2014; Silk, 2014; Willer, 2012; Willer, 2016], as well as the related work by [Bledin, 2020].

bases—that is, modal bases whose restriction is given by facts, or circumstances, in the base world that are possibly beyond the epistemic reach of any of the agents in context.[26]

The debate on information sensitivity has branched out in important ways, beyond the initial discussion of the Miners paradox. Here are a few highlights.

(i) [Willer, 2016] draws important connections between the information dependence of deontic modals and the non-monotonicity that is represented in frameworks such as update semantics.

(ii) While Kolodny and MacFarlane stopped at qualitative suppositions, [Cariani, 2016b] argues that we ought to extend the idea of information sensitivity to probabilistic suppositions—e.g. antecedents like *if the miners are likely to be in shaft A*. [Cariani, 2016b] develops a system that accommodates probabilistic information sensitivity.

(iii) Defenders of the baseline semantics have identified complex contextualist maneuvers to account for the joint acceptability of (17)-(20). The best iterations of this program [von Fintel, 2012], [Dowell, 2013], [Bronfman and Dowell, 2016] look like they can play to an empirical draw with the best information sensitive approaches. These solutions trade on an important element of the semantics that is often neglected in information sensitive systems: namely, one could represent facts about an agent's information state as included in the world that is given as input to the modal base and ordering source function.

On the other hand, one's mileage may vary on how important it is to hold on to the baseline semantics, especially considering that many of the alternatives are direct generalizations of it. It certainly seems to some theorists (including myself) that the lack of serious information dependence is not just a local empirical defect to be patched with auxiliary assumptions, but a conceptual flaw.

(iv) The idea of an information sensitive semantics can be interestingly extended to other kinds of modals. [Lassiter, 2011; Lassiter, 2017]

[26]The inadequacy of circumstantial modal bases to capture the relevant readings is discussed at some length in [Cariani *et al.*, 2013], but I have not seen work that has attempted to retrieve the explanatory goodies that the assumptions of circumstantial modal bases does in Kratzer semantics.

and [Finlay, 2014] independently discuss the case of evaluative adjectives such as *good* and its comparative *better*. Lassiter, developing a suggestion from [Levinson, 2003], also notes the nearby case of desire attributions. For further inroads in this direction, see the hot-off-the-press [Jerzak, 2019].

The linguistic territory surrounding the Miners Paradox is well charted at this point.[27] However, there may be some unexplored questions in the context of a theory of normative uncertainty.

It seems plausible to expect that a theory of normative uncertainty would include a model of how we might ascribe degrees of belief to *ought*-claims, and normative claims generally. An initial thought could be that these degrees of belief must be *probabilities*. This would extend the standard probabilist tenet that degrees of belief for ordinary factual claims ought to respect the axioms of the probability calculus. Indeed, those who don't see a principled distinction between the information conveyed by factual claims and the information conveyed by normative claims almost certainly have to accept this extension.

However accepting the extension puts us in dangerous territory because of its proximity to Lewis-style triviality (see [Charlow, unpublished] for an articulation of how triviality results might extend to deontic claims). Without getting too deep into a presentation of triviality, we can use the Miners paradox to highlight that there are parallel intuitions about credences. If these intuitions are taken at face value, they collectively give rise to probabilistically incoherent credences. Consider these questions, again on the background of the miners' scenario:

- what credence should one assign to the claim that we ought to block shaft A given that the miners are in A?
- what credence should one assign to the claim that we ought to block shaft B given that the miners are in B?
- what credence should one assign to the claim that we ought to block neither shaft?

[27] In addition to the questions about information sensitivity, the Miners Paradox has also spawned an important discussion on the philosophy of semantics. The question at the center of this thread of research is to what extent we are allowed to incorporate substantive ethical assumptions within deontic semantics. Neutralists have suggested that deontic semantics ought to be as free as possible of substantive ethical assumptions. They include [Cariani, 2014; Cariani, 2016b; Carr, 2015; Charlow, 2016; Charlow, 2018]. For an approach that more boldly incorporates substantive assumptions, see [Lassiter, 2016] and [Lassiter, 2017, ch.8].

Under the assumptions in the scenario, it seems plausible to answer "high", "high" and "high". Specifically, we might be tempted to go for the following constraints:

$$c(\bigcirc(blA|inA)) > .5$$

$$c(\bigcirc(blB|inB)) > .5$$

$$c(\bigcirc(blN)) > .5$$

However, these constraints are probabilistically incoherent.[28] After all, the epistemic space is exhausted by inA and inB. In light of that, if blN has low probability conditional on each of inA and inB, it has to have low *unconditional* probability. This means that, if the premises of this argument are accepted, a theory of normative credence must be based on principles that are not classical principles of probability. This is not the place to expand on these thoughts, except to express the hope for a clearer integration of a theory of normative uncertainty with deontic semantics.[29]

4 Puzzles of normality

Some of the classical puzzles of modal logic stem from the assumption that concepts of obligation are best modeled by normal modal operators. In particular, they stem from the assumption that *ought* expresses a concept of necessity. Philosophers of language have also reclaimed these puzzles as theoretical and empirical constraints in support of a variety of sophisticated theories. The two most important principles in this connection are:

INHERITANCE. $\mathsf{P} \vdash \mathsf{Q} \Rightarrow \bigcirc\mathsf{P} \vdash \bigcirc\mathsf{Q}$

AGGLOMERATION $\bigcirc(\mathsf{P}) \ \& \ \bigcirc(\mathsf{Q}) \vdash \bigcirc(\mathsf{P} \ \& \ \mathsf{Q})$

These principles are elementary consequences of the principles that constitute normal modal systems—the rules of necessitation and substitution, together with axiom K for $\bigcirc$.

AXIOM K $\vdash (\bigcirc(\mathsf{P}) \ \& \ \bigcirc(\mathsf{P} \supset \mathsf{Q})) \supset \bigcirc(\mathsf{Q})$

[28] To be more precise, they are incoherent relative to obvious background assumptions, such as the claim that blA, blB and blN are incompatible.

[29] For some related development of a non-factual theory of graded modal judgment (though not one that is especially focused on the case of deontic modality), see [Charlow, 2020].

We have already encountered an argument that implicitly targets normality. Specifically, in discussing Chisholm's paradox, I have taken note of Arregui's counterexample against DEONTIC DETACHMENT. This was the argument from (13)-(14) to (15). Suppose now that DD is best formalized as follows (for a conditional connective $\rightarrow$ validating *modus ponens*):

REFORMULATED DD $\bigcirc(\mathsf{P}), \bigcirc(\mathsf{P} \rightarrow \mathsf{Q}) \vdash \bigcirc(\mathsf{Q})$

The combination of AGGLOMERATION and INHERITANCE entails this.[30] So any attempt at a counterexample to REFORMULATED DD must immediately be a counterexample to one of the two principles of normality.[31]

The literature also features more direct attacks to normality. Those attacks are what we focus on in the next two sections.

4.1 Inheritance

Broadly speaking, there are two kinds of attempted counterexamples against INHERITANCE. The first consists of potential counterexamples against the entailment from $\bigcirc(\mathsf{P})$ to $\bigcirc(\mathsf{P} \vee \mathsf{Q})$. The second consists of potential counterexamples against the entailment from $\bigcirc(\mathsf{P} \ \& \ \mathsf{Q})$ to $\bigcirc(\mathsf{P})$. I will present them separately and then present the main conservative responses to them.

4.1.1 Disjunction inferences

Start with the disjunction side. The motivating observation is that [Ross, 1941]'s puzzle about imperatives extends to deontic modals. In his classic paper, Ross noted that imperatives like *mail the letter* do not seem to entail imperatives like *mail the letter or burn it.* Analogously, we might worry that (21) does not seem to entail (22)

(21) You should mail the letter.

(22) You should mail the letter or burn it.

The latter invites the inference that you may burn the letter, while the former does not. It is tempting to take this effect to be pragmatic and many theorists have endorsed this approach [Föllesdal and Hilpinen,

[30] By AGGLOMERATION, $\bigcirc(\mathsf{P}), \bigcirc(\mathsf{P} \rightarrow \mathsf{Q})$ entails $\bigcirc(\mathsf{P} \ \& \ \mathsf{P} \rightarrow \mathsf{Q})$. Because we choose $\rightarrow$ so that $\mathsf{P} \ \& \ \mathsf{P} \rightarrow \mathsf{Q} \vdash \mathsf{Q}$, INHERITANCE and $\bigcirc(\mathsf{P} \ \& \ \mathsf{P} \rightarrow \mathsf{Q})$ yield $\bigcirc(\mathsf{Q})$.

[31] There is room to escape this consequence if we refuse to formulate deontic detachment as in the above, and instead go in for NARROW DEONTIC DETACHMENT from Section 3.2.

1971; Wedgwood, 2006]. For example, Wedgwood suggests accounting for the badness of (22) in terms of a Gricean quantity implicature.

> There is an obvious Gricean explanation for why (22) seems an odd thing to say. It is much less informative than something else one might say–namely (21). Asserting the weaker claim would tend to be a useful contribution to a conversation only if one was not in a position to assert the stronger claim.

However, [Cariani, 2013a, §. 6] argues that it is implausible to take it as an pragmatic implicature, since it does not seem to behave as one.[32]

As an alternative, [Cariani, 2013a] offers a framework that could deal with this kind of phenomenon semantically. The design challenge that needs to be met by any semantic framework of this kind is to explain how to account for the apparent falsity of (22) without landing on the evidently false hypothesis that $\bigcirc(\mathsf{P})$ can only be true if all the relevant P worlds are permissible. That hypothesis would lead to a vast proliferation of false *ought* judgments: consider, for instance, a true-sounding *ought* claim like *you ought to take good care of yourself*. There are plenty of overall impermissible ways taking good care of oneself, but they do not undermine the truth of the *ought*-claim.

[Cariani, 2013a] proposes that this be viewed as a form of 'coarseness' of the semantics. This is handled within a formalism that, taking a lead from [Yalcin, 2011], is referred to as *resolution semantics*. The key idea of that formalism is that the deontic orderings that the semantics accesses are not orderings of worlds but orderings of alternatives — where alternatives are coarser objects than worlds.[33]

The discussion of [Cariani, 2013a] fails to identify the complete logic of the resolution semantics for $\bigcirc$. This logical project has been taken up

[32]The dialectic interacts in significant ways with developments in the theory of scalar implicatures that are too rich to detail here. A popular approach, described for instance in [Chierchia *et al.*, 2008], treats scalar implicatures as entirely grounded in syntactic and semantic facts. The view in [Cariani, 2013a] is that if it turns out that there is a viable account of Ross-type phenomena within this framework, that would still be a win for the non-pragmatic camp, even if the theory turned out to have a different shape than initially anticipated.

[33]Some of my later work on information sensitivity [Cariani *et al.*, 2013; Cariani, 2013b; Cariani, 2016b], seeks to show that this idea has valuable applications even if one *accepts* INHERITANCE. Indeed, I maintain that this is critical to get a proper treatment of information-sensitivity. In addition, there are applications of these contrastive ideas to expressions other than modals. For example, [Snedegar, 2017], mounts an impressive case that contrastivism helps solve some puzzles in the logic and semantics of reasons-claims.

and completed by Van De Putte in a striking contribution [Van De Putte, 2018]. Van De Putte shows that the non-normal obligation operator in [Cariani, 2013a] can be decomposed in three normal operators, each of which has a well-understood logic. In addition to the intrinsic interest of Van De Putte's results, his work stands as a model of what a two-way interaction between logic and philosophy of language might look like.

4.1.2 Conjunction inferences

As for conjunction elimination, the central puzzle case is Frank Jackson's Professor Procrastinate case [Jackson, 1985; Jackson and Pargetter, 1986].[34] In Jackson's story, Prof. Procrastinate is asked to write a book review on a subject on which he is the foremost expert. However, Procrastinate's disposition is such that if he accepts the commitment to write the review, he won't write it. Jackson's judgments in this case are that *Procrastinate ought to accept the commitment and write the review* but *Procrastinate ought to accept* is false (because if he accepts he won't write).

Jackson thought that the intuitions surrounding the Procrastinate case are evidence for an *actualist* semantics, according to which $\bigcirc(\mathsf{P})$ is true at w if the value of the closest P-world to w exceeds the value of the closest $\neg\mathsf{P}$-world. This sorts of semantics can be viewed as an ancestor to Arregui's proposal. What they share is the idea that there is a counterfactual element in deontic modality.

Cariani [2009; 2013a] points out that the counterfactual element in the Procrastinate case is not necessary to get the relevant intuitions going. More specifically, we don't have to stipulate that Procrastinate *won't* write the review. All that is necessary is the stipulation that it is *very unlikely* that he will. Partly with this kind of puzzle in mind, [Cariani, 2009] and, much more extensively, [Lassiter, 2011; Lassiter, 2017] developed decision-theoretic accounts of the meaning of *ought*.[35] According to these decision-theoretic approaches, Procrastinate ought to accept and write the review because the expected value of accepting

[34] [von Fintel, 2012] notes that these cases have antecedents in a very similar case by [Kamp, 1973, pp.59-60] and are importantly like some cases that have been described in the literature on desire ascriptions.

[35] For an initial presentation of a decision-theoretic semantic within the deontic logic literature, see [Goble, 1993] (though Goble does not apply the semantics to Procrastinate cases). For a very different kind of broadly decision-theoretic account—one which actually validates INHERITANCE—see [Wedgwood, 2016]. For some reasons to avoid this decision-theoretic semantics, see the literature on semantic neutrality, including [Cariani, 2014; Cariani, 2016a; Carr, 2015; Charlow, 2016; Charlow, 2018].

and writing is higher than the appropriate threshold. (The threshold itself might be understood contrastively as the expected value of the salient alternatives). By contrast, merely accepting is low in expected value, given the low probability of the good outcome if Procrastinate accepts.

That said, rejecting INHERITANCE in Procrastinate cases—even these probabilized ones—is not the sole province of expected utility accounts. Both the resolution semantics of [Cariani, 2013b] and Arregui's semantics for Chisholm's paradox make the same prediction about these cases, without appealing to expected utilities.

4.1.3 Responses and Arguments in Favor of Monotonicity

There are important criticisms for all of these moves by defenders of INHERITANCE. [von Fintel, 2012] objects that the disjunction in Ross's Puzzle is free choice disjunction[36] and that the intuitions in Procrastinate cases can be addressed as involving context shift, and specifically as involving expansions of domains of quantification as more possibilities are made salient. We reject *Procrastinate ought to accept* when the possibility of writing is not salient. We accept *Procrastinate ought to accept and write* because that ideal possibility needs to be salient for this to be even evaluable. ([Bronfman and Dowell, 2018] also adduce similar context-shift considerations.) Von Fintel notes one consideration that supports the context-shift move. It is extremely bad to conjoin the two Procrastinate sentences. For example, (23) feels like a contradiction.

(23) # Procrastinate ought to accept and write, but he ought not to accept.

Von Fintel is right that any viable theory founded on the rejection of INHERITANCE must account for why these sentences sound contradictory. Little has been said by critics of INHERITANCE, myself included, to account for such judgments in a systematic way.[37]

[36] As I have noted, von Fintel suggests treating Ross's puzzle as free choice. This possibility was anticipated in [Cariani, 2013a, p. 551]. The response provided there was that free choice arises equally with deontic and non-deontic interpretations of modals, but Ross-type phenomena don't. Intuitions about failures of INHERITANCE are much weaker for epistemic *must*, than they are for even deontic *must*. In fact, I suspect those intuitions can entirely be accounted for in pragmatic terms.

[37] Von Fintel also notes another debt that such theories incur. There is a rich linguistic literature on the status of NPI's (Negative Polarity Items). These are items like *any* that are only allowed in special environments. The standard view is that what makes an enviroment "special" in the relevant sense is that it is downward

Recent work has also elaborated on the connection between Ross's puzzle and free choice phenomena, though perhaps not quite in the direction envisaged by [von Fintel, 2012].[38] In particular, substantial progress was made in work by Fusco (specifically [Fusco, 2015; Fusco, 2014]) who provides an analysis of permission in terms of ratifiability and a two dimensional semantics for disjunction that allows a unified account of Ross-type phenomena and free choice permission. Here is a quick sketch of Fusco's semantics. Points of evaluations are triples $\langle s, y, x\rangle$ consisting of a state s, the actual world y (unshiftable by modal operators), and the world of evaluation x. First, define an auxiliary concept Alt_w:

$$Alt_w(\mathsf{P}, \mathsf{Q}) = \{\mathsf{R} \in \{\mathsf{P}, \mathsf{Q}\} \mid \mathsf{R} \text{ is true at } w\}$$

Informally, this returns whichever subset of $\{\mathsf{P}, \mathsf{Q}\}$ contains all and only the sentences that are true at w. Letting R be a deontic accessibility relation, say:

$$[\![\mathsf{P} \text{ or } \mathsf{Q}]\!]^{s,y,x} = 1 \text{ iff } \exists \mathsf{R} \in Alt_y(\mathsf{P}, \mathsf{Q}), [\![\mathsf{R}]\!]^{s,y,x} = 1$$
$$[\![\bigcirc(\mathsf{P})]\!]^{s,y,x} = 1 \text{ iff for all worlds } v \in s \text{ s.t. } xRv, [\![\mathsf{P}]\!]^{s,y,v} = 1$$

Informally, a disjunction is true (relative to s, y and x) if one of the alternatives drawn from P and Q that hold at y is true (again, relative to s, y, and x). The semantics for $\bigcirc$ is relatively standard (except for the rich points of evaluation).

monotonic (Here $\Phi(\_\_\_)$ is downward monotonic iff $\Phi(\mathsf{Q}) \models \Phi(\mathsf{P})$ whenever P entails Q.) The problem is that *any* is licensed in *You don't have to bring any alcohol to the party.* This is captured by the INHERITANCE validating semantics because $\neg \bigcirc (\cdot)$ is downward entailing iff $\bigcirc(\cdot)$ is upward entailing—that is iff $\bigcirc$ satisfies INHERITANCE.

I concede to von Fintel that many of the INHERITANCE-rejecting proposals don't do a good job of tackling this issue, and I won't do much better here. With that said, I am more optimistic about the prospects of addressing this concern. For one thing *any* is licensed in environments that are clearly *not* downward monotonic. One example is *It's fifty percent likely that Mary didn't eat any of the cookies.* So what really seems to matter is that the environment be in some sense *locally* downward entailing. That requires getting clear about what it is to be "locally" downward entailing. Depending on how *that* story goes it seems possible to reconstruct the semantics of *ought* so that it combines a non-monotonic component and a monotonic one.

[38] I lack the space for a summary of the free choice literature here, which goes well beyond the modal case. But *that* is also relevant to the theme "Deontic Logic and Natural Language". So, it is worthwhile identifying some of the key references. [Kamp, 1973] is justly heralded as a classic. Among recent works that are of specific importance to the study of deontic concepts are: [Aher, 2012; Barker, 2010; Starr, 2016; Willer, 2018].

Let us follow along with this semantics' evaluation of (22)—*you ought to mail the letter or burn it.*

$$[\![\bigcirc(m \text{ or } b)]\!]^{s,y,x} = 1 \text{ iff for all worlds } v \in s \text{ s.t. } xRv, [\![m \text{ or } b]\!]^{s,y,v} = 1$$

$$[\![m \text{ or } b]\!]^{s,y,v} = 1 \text{ iff } \exists \mathsf{R} \in Alt_y(m, b), [\![\mathsf{R}]\!]^{s,y,x} = 1$$

Suppose s contains worlds in which you burn the letter without mailing it (not that it would be helpful to mail a charred letter). Suppose further one such a world is actual. Then the disjunction m or b will collapse on b, and so $\bigcirc(m \text{ or } b)$ will collapse on $\bigcirc(b)$ which can easily be made false even if $\bigcirc(m)$ is true. The story can be completed with a "ratifiability"-based account of permission that also delivers an account of free choice.

Fusco's view shares with the accounts by Jackson and Arregui discussed above the idea that what ought to be the case might depend in part on what actually is the case. However, in Fusco's system this effects stems crucially from the semantics for disjunction, and it is not especially linked to the semantics for obligation. This means in particular that her semantics treats Ross type phenomena as fundamentally different from apparent failures of conjunction elimination inside $\bigcirc$. Furthermore, for that very reason, the contingency that Fusco recognizes in deontic claims is unlike the contingency we saw at work in Chisholm's paradox.

4.2 Agglomeration

Next up is the delicate matter of the agglomeration principle.

AGGLOMERATION $\bigcirc(\mathsf{P}) \ \& \ \bigcirc(\mathsf{Q}) \vdash \bigcirc(\mathsf{P} \ \& \ \mathsf{Q})$

There are different putative counterexamples to AGGLOMERATION, and it seems likely that these ought to be evaluated independently.

Deontic conflict

The most famous examples involve deontic conflicts. Simple, illustratives examples of deontic conflict involve moral dilemmas (but note that there are many deontic conflicts that are not dilemmas). Maybe all of the following are true: (i) I ought to travel to a different country to perform my civic duties, (ii) I ought to stay home to attend to my ailing mother but (iii) it's not the case that I ought to do both. The study of logics for normative conflicts is a remarkably well developed area of investigation in deontic logic (see [Goble, 2013] for a survey).

My (possibly controversial) opinion is that the philosophy of language literature and linguistics hasn't produced much that is new on deontic conflict. Most typically, theorists recognize the need for an account of deontic conflict, but relegate that account to an additional module to be separately injected in one's semantic framework.[39] This is not to say that the issue gets no substantive discussion. [Lassiter, 2011] sought to reduce dilemmas to a more general pattern of non-agglomerating *oughts*. But, as Lassiter himself recognizes in later work [Lassiter, 2017, §8.11], this cannot be exactly right. There is something distinctive going on in the case of dilemmas—something that requires additional treatment even if, as Lassiter does, we reject AGGLOMERATION for independent reasons.

Indeed [Lassiter, 2017, §8.11] is one of the few exceptions to my claim that new ideas in deontic dilemmas from the linguistics side are few and far between. In that section of *Graded Modality*, Lassiter explores the idea that there might be connections between the structural features that give rise to deontic conflicts and the structure of multi-dimensional adjectives. You might think for instance that *clever* tracks many dimensions of cleverness, and similarly you might think that deontic words (or, better, priority modal expressions generally) track many dimensions of priority. This is an interesting idea that, though still in the early stages, is likely to draw attention and development in the future. [40]

Other challenges to agglomeration.

Some authors [Jackson, 1985; Finlay, 2014; Lassiter, 2011] propose accounts of deontic *ought* that involve more extensive violations of AGGLOMERATION. That is to say, violations of AGGLOMERATION that go beyond the isolated case of deontic conflict. For example, Jackson proposes this kind of example:

> Attila and Genghis are driving their chariots towards each other. If neither swerves, there will be a collision; if both swerve, there will be a worse collision [...]; but if one swerves and the other does not, there will be no collision. Moreover if one swerves, the other will not because neither wants a col-

[39] Versions of this strategy are gestured at in [Cariani, 2013a] and [von Fintel, 2012].

[40] There is interesting and related work on how to extract deontic domains from inconsistent premise sets, e.g. [Silk, 2017]. There is also important work concerning how deontic conflicts might figure in a theory of *reasoning*: see [Nair, 2014; Nair, 2016].

> lision. Unfortunately, it is also true to an even greater extent that neither wants to be 'chicken'; as a result what actually happens is that neither swerves and there is a collision. It ought to be that Attila swerves, for then there would be no collision. [...] Equally it ought to be that Genghis swerves. But it ought not to be that both swerve, for then we get a worse collision. [Jackson, 1985, p.189]

[Cariani, 2016a, p.400-401] expresses skepticism about treating this as a general counterexample to AGGLOMERATION.[41] But let us concede Jackson's description of the data, for the sake of presentation. Note that the example, in its intended reading, does not involve multiple conflicting conflicts of value. Jackson's *ought*-claims are understood as relative to a world-evaluation method according to which the best possible outcome is just to avoid collisions, and in which the worse the collision the worse the world. Given all that, if this is a counterexample to AGGLOMERATION, then there are counterexamples to AGGLOMERATION that do not arise from deontic conflicts.

Corresponding to these putative violations are semantic theories that block these instances of AGGLOMERATION. For Jackson, α ought to fly is true at world w if the nearest world in which α flies is better than the nearest world in which α does not fly. Agglomeration fails when the nearest P & Q-world is bad, but the nearest P-world is good (because it is a P & $\neg\mathsf{Q}$-world) and the nearest Q-world is also good (because it is a $\neg\mathsf{P}$ & Q-world). For Lassiter, α ought to fly is true relative to a probability p and utility u function if the expected utility of flying (calculated on the basis of p and u) exceeds the expected utility of the alternatives.

Cariani [2016a] argues that the path traced by these proposals ends up in a problematic place. In particular, nearly every implementation of this idea ends up violating something much more plausible than AGGLOMERATION — namely the inference:

WEAKENING. $\bigcirc(\mathsf{P}), \bigcirc(\mathsf{Q}) \vdash \bigcirc(\mathsf{P} \vee \mathsf{Q})$.

The few combinations that avoid this pitfall end up validating AGGLOMERATION, either in general or in Jackson's alleged counterexample that

[41] I don't doubt that there are strong judgments in the direction of Jackson's intuition. In fact, the naked effect is strongly supported by an experimental study by [Lassiter, 2017, §8.7]. I doubt that Jackson's chicken case is well understood as a case in which all the contextual parameters are held fixed throughout the evaluation of the argument. I have similar views about the scenario in Lassiter's experiment. For a fuller articulation of this kind of diagnosis, see [Boylan, ms].

was supposed to be a key motivating factor.

In response, [Lassiter, 2017] makes two interesting suggestions: one idea (§8.10) is to stipulate away the counter-models. The move is not entirely unprincipled. After all, Lassiter notes that the countermodels described in [Cariani, 2016a] are merely abstract: they show that it is possible for the semantics to invalidate weakening given certain parameter values, without describing cases that would instantiate that structure. So, perhaps there is room to rule them out by excluding the appropriate parameter values as inadmissible. If it is true that the abstract countermodels are hard to connect to intuitive scenarios, the theory incurs no *predictive* cost. The other approach Lassiter explores is to take WEAKENING as invalid, but recover it via scalar mechanisms. For elaboration of this alternative approach, see §8.14 of [Lassiter, 2017].

The last recent development I want to highlight here is Boylan's ([Boylan, ms]) argument that *epistemic* but not *deontic* oughts fail to agglomerate. I won't present Boylan's counterexamples to epistemic *ought* AGGLOMERATION here, so as to not steal his thunder. But if he is right, and if we want to give a unified semantics for epistemic and deontic *ought*, we will need to reach for an abstract semantics that invalidates AGGLOMERATION and then recover deontic AGGLOMERATION *via* some special claim in the theory of flavors. Boylan's paper develops one way to do this.

5 Varieties of deontic strength

Another important research thread concerns the varieties of deontic strength. Theorists frequently associate English *ought* and *must* with distinct concepts of obligation. The *ought* concepts are generally taken to be weaker than the related *must*-concepts. Let us start, then, by formulating that hypothesis officially:

STRENGTH ASYMMETRY. unembedded *must*-claims are logically stronger than their *ought*-counterparts

This difference is sometimes represented in terms of the claim that *must* is a 'strong necessity modal' while *ought* is a 'weak necessity modal'. Because this formulation ignores the views that deny that *ought* is any kind of necessity modal, it is better to start by formulating the idea in more general terms.

The STRENGTH ASYMMETRY thesis has been discussed in the deontic logic literature, well before its appearance in natural language semantics [Sloman, 1970; McNamara, 2010]. However, linguistic work (largely

spawned by [Copley, 2006] and [von Fintel and Iatridou, 2008]) has provided additional theoretical and empirical depth to this discussion.

5.1 Motivating Data

Below is a simple paradigm, drawing on von Fintel and Iatridou's examples:

(24) You ought to wash your hands but you don't have to.

(25) #You must wash your hands but you don't have to.

(26) #You must (/have to) wash your hands, but it's not the case that you ought.

The perceived inconsistency of (25) suggests that *must* and *have to* have similar levels of strength. By contrast, the consistency of (24) suggests that *ought* is weaker than both *must* and *have to*. A moment's reflection on similar data reveals that *should* patterns with *ought* — in the sense that replacing *ought* with *should* in (24) and (26) generates analogous judgments.

These observations invite the view that there are two levels of deontic force. As postulated in STRENGTH ASYMMETRY, the stronger level is occupied by *must* and *have to*, and the weaker level is occupied by *ought* and *should*.

Before moving to the central theoretical questions, let us dwell on one more empirical aspect of the asymmetry. In light of STRENGTH ASYMMETRY, we may consider how permission claims pattern. In particular, data resembling the paradigm from von Fintel and Iatridou suggests that English *may* is not the dual of *ought*. Indeed, if *may* has to be the dual of something, *must* seem to be the right choice.

(27) You ought to wash your hands, but you may not.
$[ought(\mathsf{P}) \ \& \ \neg may(\mathsf{P})]$

(28) # You must (/have to) wash your hands, but you may not.
$[must(\mathsf{P}) \ \& \ \neg may(\mathsf{P})]$

These considerations open the interesting data question whether *ought* has a dual in English (and indeed whether in other languages, the appropriate translations of *ought* happen to have dual).

This question has received substantially less attention than the STRENGTH ASYMMETRY, but it was recently addressed by [Beddor, 2017]. Beddor uses the term "faultlessness" to express the dual of the

concept of weak necessity (weak necessity being what he takes *ought* and *should* to denote). The linguistic question is whether faultlessness is linguistically realized, in English or other languages. Beddor goes on to propose that faultlessness is expressed in English by the phrase *is justified*, as it occurs in sentences like

(29) Ada isn't justified in attempting a cartwheel.

Partial support for this idea seems to come from the fact that (29) sounds roughly equivalent to *Ada should not attempt a cartwheel*.

We could check this by inspecting consistency judgments. The following are two key test cases for Beddor's conjecture.

(30) #Ada should attempt a cartwheel but she isn't justified in doing so.

(31) #?Ada is justified in attempting a cartwheel but she should not.

My own judgment is that (30) is unfixably bad, while (31) might be repaired by having the modal pick on something other than the initial set of reasons that provide the justification. This seems to accord with Beddor's hypothesis. If Beddor's hypothesis is correct, we would have the surprising consequence that some grammatical modals of English have lexical modals as their duals.

5.2 Accounts of the strength asymmetry.

The central theoretical question concerning STRENGTH ASYMMETRY is what is the correct semantic explanation of the phenomenon. Recent literature has explored a variety of approaches. We will consider two types of approaches, and then mention an approach that does not quite fit either type. These are the two types:

> TYPE 1. *must* and *ought* are analyzed as quantifiers over worlds but associated with different domain-generation rules.
>
> TYPE 2. *must* and *ought* are degree expressions that are sensitive to different thresholds.

Implementations of these ideas are typically elaborated within very specific semantic frameworks. However, the broad ideas are sufficiently modular to allow presentation independently of those specific implementations. I refer the reader to the primary sources referenced below for detailed implementation.

The domain-centric approaches (TYPE 1) all involve the classical assumption that *ought*, *must* and their cognates are necessity operators of some sort. In this context, it make senses to refer to the former as 'weak necessity operators' and to the latter as 'strong necessity operators'. The central tenet of TYPE 1 theories is that domains of the weak operators must be subsets of the domains of the strong ones.[42]

Orientation check: if $D \subseteq D^+$, and $\Box$ and $\Box^+$ quantify respectively over D and D^+, then $\Box^+$ will be at least as strong as $\Box$. Indeed, under basic assumptions, it will be strictly stronger whenever the inclusion between domains is itself proper.

As von Fintel and Iatridou put the point:

> Our conception of weak necessity then makes them universal/necessity modals just as much as strong necessity modals are. What makes them weaker semantically is that they have a smaller domain of quantification: strong necessity modals say that the prejacent is true in all of the favored worlds, while weak necessity modals say that the prejacent is true in all of the very best (by some additional measure) among the favored worlds. [von Fintel and Iatridou, 2008, p.119]

The domains of quantification are generated by two layers of ordering sources. The *primary ordering source* contains those considerations that are hard requirements, while the *secondary ordering source* contains those considerations that are, in some sense, "additional criteria" that are significant for ranking worlds but not strictly required.

It is sometimes objected that this view does not explain the contrast. What it does, the objection goes, is merely writing the contrast into the semantics. More specifically one may press questions such as:

- How do we understand the talk of "hard requirements"?
- And what makes a criterion "additional"?

One attempt to address this challenge is found in [Chrisman, 2012]. Chrisman proposes that the key distinction is between the *requirements* and the *recommendations* of morality. Of course, since *ought* and *must* may be used well beyond the domain of moral discourse, this distinction would have to be suitably generalized to cover all priority modals. In

[42] More precisely, if the domains vary from world to world (as they do in the baseline Kratzerian theory) the domains of the weak operators *at each world of evaluation* w are required be subsets of the domains of the strong ones at w.

particular, for every source of priorities, we should be able to distinguish between what it requires and what it recommends.

Another substantive attempt to articulate a distinction in this neighborhood is [Rubinstein, 2012]. [Rubinstein, 2012] proposes that *must* and *ought* signal different levels of commitment to the priorities that ground them. In particular, strong necessity modals track those priorities that are commonly accepted as such by conversational participants, while weak ones additionally track those items that are being proposed as additions to a to-do list, but are generally held to be up-for-grabs.

Let us move on to TYPE 2 approaches— those that propose that deontic modals are threshold expressions. These approaches start off by rejecting the idea that *ought* and *must* are necessity operators of any kind. Instead, they lean on an extended analogy between these modals and the probability operator *likely*. We could model *likely*(P) as claiming that P has probability greater than some threshold (perhaps .5). Similarly, we could model *ought*(P) as saying that P has a degree of 'oughtiness' greater than some threshold. Needless to say, an important sub-task consists in explaining what these "degrees of oughtiness" are. According to these views, we might model the difference between *ought* and *must* roughly on a treatment of threshold adjectives that are on the same scale but point to different threshold points (say *hot* vs. *warm*). For example, Lassiter [2011; 2017] proposes that both *ought* and *must* track expected values, but also that *must* demands higher thresholds than *ought*.

An alternative proposal in this family is found in [Finlay, 2016]. For him, *ought* P means that P is more likely, conditional on the agent's contextually salient ends, than the alternatives. By contrast, *must* P means that P is certain, conditional again on the agent's contextually salient ends being realized. The asymmetric entailment is then recaptured by the simple fact that the truth-condition for *must* requires the truth-condition for *ought* to be satisfied.

These proposals generally do well at accounting for the core data supporting STRENGTH ASYMMETRY. (Some, but not all, TYPE 2 approaches have trouble with the permission data in (28), depending on how permission is analyzed.) Insofar as they run into trouble, it is because they run afoul of some other desiderata (see for example the discussion of AGGLOMERATION violation in Section 4.2 above).

The last account I want to highlight does not neatly fit either the TYPE 1 or the TYPE 2 mold. According to [Silk, forthcoming], the fundamental difference between *must* and *ought* is that the former, but not the latter, requires that its prejacent be a necessity with respect to

the actual priorities. Informally, *must* P says that P is necessary in the *actual* world; *ought* P says that P is necessary in a range of worlds that are relevantly related to the actual world but needn't *be* the actual world. There are two things that make this theory interesting: for one thing, it features a new way of capturing the "counterfactuality" of *ought*; for another, Silk is after the idea that there appear to be more levels of uncertainty about *ought* claims than there are about *must* claims. The most striking way in which Silk's account is at variance with the rest of the literature is that on his view *must* P does not entail *ought* P—so that *ought* and *must* are logically independent of each other. Silk gives some preliminary arguments (§4) that this is a feature and not a bug of his proposal (but the matter probably deserves greater scrutiny).

6 The grammar of action

It is common to assume that *ought*, *must* and *may* are propositional operators. The more carefully we integrate our analysis with linguistics, the more it is worth scrutinizing this assumption. An important development in this direction concerns the proper modeling of the interaction between agency and obligation in the context of a linguistically plausible model of deontic modality.

So here is some old news. Consider:

(32) Gaia should dance with Iris.

This is widely believed to have multiple interpretations, even fixing a broadly deontic flavor. On one interpretation, it identifies the ideal state by whatever salient criteria. In particular, it claims that the ideal state is one in which Gaia and Iris dance together. On another interpretation, it conveys that same content *plus* something extra—something to the effect that it is *up to* Gaia to make sure that she dances with Iris.

To elicit the first interpretation, consider a context in which the organizers of a ball are choosing which people are to dance with each other. It may well be best for the organizers that Gaia and Iris dance together, but it doesn't behoove Gaia to bring this about. To elicit the second interpretation, consider instead a context in which Gaia herself is choosing a dance partner. It is standard to label these the *ought-to-be* interpretation and the *ought-to-do* interpretation [Feldman, 1986]. Though it is standard, however, this is not the only way of conceptualizing the dichotomy. Some authors distinguish between *agentive* interpretations and *non-agentive* interpretations of modals. Others contrast *deliberative*

interpretations with *non-deliberative*, or *evaluative*, ones. The plurality of terminological choices suggests that it is probably a mistake to think of this as *one* dichotomy.

6.1 Logical approaches

There are many ways to design a formal system that is capable of generating a distinction like the one we are after. An unimaginative approach might stary by characterizing two sentential operators $\bigcirc_{be}$ and $\bigcirc_{do}$. Immediately, we might worry that this treatment does not attempt to explain the plurality of interpretations of (32). At best, the approach reflects some prior understanding of what the distinction might be.

What weight to give to this worry depends on what kind of application we have in mind. There are applications for which the multiple operator approach is entirely unobjectionable. After all, logic alone does not demand explanations of empirical phenomena. Moreover, a theorist might just be interested in questions like: what is the logic of *ought-to-be* operators?

Having said that, one doesn't need to be deeply invested with linguistics to ask more than what the two-operators approach can provide. For instance, one might want to design a system that can capture, without simply stipulating them, the logical interactions between deontic claims and sentences ascribing agency. This perspective is found in the familiar *stit* framework, as developed for instance in [Horty and Belnap, 1995; Belnap *et al.*, 2001; Horty, 2001]. In this framework, we add a family of sentential operators $stit_{\alpha}(\cdot)$ with the intended interpretation that agent α brings about the state of affairs that is described by the input proposition. So, $stit_{\alpha}(rain)$ might be a formalization of:

(33) α sees to it that it rains.

I will not develop a full *stit* framework since it is likely extremely familiar to the readers of this handbook. The key idea for our purposes is that *stit* framework allows us to capture the distinction between *ought-to-be* and *ought-to-do* in terms of a single obligation operator $\bigcirc$. This could be achieved by feeding $\bigcirc$ inputs that differ in their agentivity. So, $\bigcirc$(Gaia dance with Iris) symbolizes that it ought to be the case that Gaia dances with Iris. By contrast, $\bigcirc(stit_{Gaia}$(Gaia dance with Iris)) symbolizes that Gaia ought to see to it that she dances with Iris.

6.2 Linguistic concerns

Once again, this is unobjectionable logical development. But, for some applications, we need a more direct account of the connection between a formalism for *oughts* and agency and natural language.

In an influential paper [Schroeder, 2010] has charted a set of challenges for the *stit* analysis interpreted as a piece of philosophy of language.[43] As I interpret him, Schroeder argues that there are two linguistic constraints that are violated by the *stit* analysis. For future reference, let me give these constraint some names:

ADHERENCE. An *adherent* theory must predict, without overgenerating, which interpretations of any deontic sentence is available in any given context.

FAITHFULNESS. A *faithful* theory must respect, and ideally *account for*, relevant generalizations emerging from natural language syntax and semantics.

Schroeder argues that versions of the *stit* view generally violate both these desiderata.

The *stit* view fails to be adherent predicting agentive interpretations that are, in fact, not available. To see the force of the argument, we must first expand the *stit* view. Of course, we never say things like,

(34) It ought to be that Lisa sees to it that she runs.

We more often say things like:

(35) Lisa ought to run.

For the *stit* view to account for the agentive reading of (35), it needs to take on two additional claims. First, that a *stit*-operator is present but unpronounced in (35). Second, that the other implicit arguments that are required to fill out the logical form appropriately can also be reconstructed, even if they are not explicitly provided. In other words, (35) needs a logical structure roughly like (38). (Note: in (38) I use small caps for unpronounced material and parentheses to mark scope)

(36) $\bigcirc$ ($\text{STIT}_{\text{LISA}}$ (Lisa run))

Finally, one must hypothesize that the new, unpronounced occurrence of LISA is somehow dependent on the overt subject of (35). This avoids the

[43] In fairness this is not the spirit in which the analysis is typically proposed.

potential worry that (35) might end up being assigned the meaning that it is up to Simone to see to it that Lisa runs.[44] Call this expanded version of the *stit view* "the agency-in-the-prejacent" view" (AIP for short).

The problem Schroeder identifies is that AIP appears to overgenerate. Imagine a conversation in which two people emphatize with a friend, Luckless Larry, who has a had to endure a remarkable series of misfortunes. Suppose that one of them says:

(37) Larry ought to win the lottery.

According to Schroeder, (37) does not have an agentive interpretation—not, at least in contexts in which it is common ground that Larry has no ability to affect the lottery.[45] But it is unclear how AIP avoids this consequence. After all, the same mechanisms that generate (38) should generate a logical form roughly like:

(38) $\bigcirc$ ($\text{STIT}_{\text{LARRY}}$ (Larry wins the lottery))

And that logical form should be assigned an agentive reading.

In addition—and perhaps more worryingly—Schroeder argues that AIP also violates FAITHFULNESS. Let us go back to the transition from (35) to (38). What must justify this transition is that *ought* is assimilated to a *raising* verb.[46] Without getting into the details of the syntax of raising constructions, the claim is that the relation between (38) and (34) parallels the relation between (39) and (40).

(39) Lisa appears to be tired.

(40) It appears that Lisa is tired.

In particular, the overt subject of (38) does not correspond to an argument of *appears*. Instead, it is in fact the subject of the "lower" verb, raised to the apparent position of subject of "higher" one. From the point of view of semantics, this means for instance that the semantic engine first evaluates the composition of *Lisa* and *be tired* and then applies the sentential operator *appears*.

[44] More specifically, the worry is that it might otherwise get the logical structure $\bigcirc$ ($\text{STIT}_{\text{SIMONE}}$ (Lisa run)).

[45] [Bronfman and Dowell, 2018] push back on these judgments. According to them, the interpretation is not ruled out by the grammar, and instead it is simply made far fetched by contextual knowledge.

[46] The category of *raising* verbs is from syntax where it is contrasted with the category of *control* verbs. Schroeder relies on textbook presentation from [Radford, 2004], which is as good as any reference.

Schroeder's point is that *ought* does not unambiguously meet the standard criteria for raising verbs. Contrast:

(41) Inez should examine June.

(42) June should be examined by Inez.

The hypothesis that *should* is a raising verb predicts that these are equivalent. This is why: suppose *should* were a raising verb, then the logical forms of (41)-(42) are roughly as in (43) and (44) respectively.

(43) ◯(Inez examine June)

(44) ◯(June be examined by Inez)

These can only have different truth-conditions if their prejacents have different semantic values. But their prejacents are related by passivization and it is a default assumption that passivization preserves semantic value—i.e., that sentences that are related by passivizations make the same contributions to the meanings of complex expressions that embed them. So, if *ought* is raising, (41)-(42) must have the same interpretations. a

Schroeder argues there seems to be a sense in which this is true, but that sense is not the one that is associated with agentive interpretations of *ought*. Instead, that alternative sense seems closer to the *ought-to-be* interpretations. (To be exact, Schroeder avoids the *do/be* dichotomy and distinguishes between a *deliberative* and *evaluative* interpretation of *ought*.) One way to see this is that (41) and (42) in fact have *different* agentive interpretations. In its agentive interpretation, (41) demands that Inez see to it that she examine June. By contrast, (42) demands that June see to it that she is examined by Inez. So the agentive interpretations are not synonymous, and hence (41)-(42) cannot be raising verbs.

Schroeder's proposal is that *ought* is lexically ambiguous: one item, the "evaluative" *ought* is a sentential operator and classified as a raising verb; the other item, the "deliberative" *ought*, takes two separate arguments, an agent and an action.[47] In the possible worlds framework (which Schroeder does not adopt), the evaluative *ought* denotes a function from sets of possible worlds to truth-values.[48]

[47] There are quite a few antecedents for this kind of proposal in both deontic logic and linguistics. See, among others, [Thomason, 1981a; Brennan, 1993], and [Portner, 2009, §4.1.4 and §4.3.2].

[48] In the possible worlds framework (which Schroeder does not adopt), these can be assigned to types. Let t= the type of truth-values; e= the type of individuals

Schroeder's arguments have attracted a variety of responses. [Chrisman, 2012] argues that there are some diagnostics for control that are not satisfied by purportedly deliberative *oughts*. In particular, he focuses on the fact that typical control predicates can be nominalized. From *begin*, we can form *beginner*; from *want* we can (with some strain) form *wanter*; but it is ungrammatical to nominalize *ought*—there are no *oughters*. It is however unclear exactly how significant this observation is, since we already know that English *ought* is independently known to accept limited morphological combination (for example it lacks infinitive forms).

Other responses attempt to provide accounts of Schroeder's data that are nonetheless compatible with an ambiguity theory. [Finlay and Snedegar, 2014] propose a *contrastivist* account of the motivating data. For them, it is important to recognize that *ought*-sentences are relative to alternatives. Additionally, these alternatives are connected with focal stress. That is to say, that sentences like *Gaia ought to dance with Iris* suggests different things depending what speakers stress. Stressing *Gaia* suggests that the alternatives are possible dance partners for *Iris*. Stressing *dance* suggests that the alternatives are possible activities for Gaia. Stressing *Iris* suggests that the alternatives are possible dance partners for *Gaia*. The key claim is that in order to get a deliberative/agentive/ought-to-do reading, the alternatives (generated by this or other similar means) must be in some sense "available" to the agent. This helps explain why we cannot get a deliberative interpretation of *Larry ought to win the lottery*. After all any such interpretation would have to include an unavailable option because winning the lottery is not (except in some exceptional cases) an available option.

[Bronfman and Dowell, 2018] defend a thoroughly contextualist approach by carefully working through Schroeder's arguments against ADHERENCE. Bronfman and Dowell show that the linguistic data underpinning Schroeder's argument are a little more sensitive to contextual variations than Schroeder initially suggested. Their theoretical proposal is to build on a relatively orthodox version of the baseline semantics. Their main contribution consists in articulating just how flexible the contextual apparatus is.[49]

and s=the type of possible worlds. Then the types of the two *oughts* are respectively $\langle\langle s,t\rangle,t\rangle$ and $\langle e,\langle e,\langle st\rangle\rangle,t\rangle$.

[49] To be clear, one could replicate their general strategy within many other semantic frameworks. Bronfman and Dowell present this as a defense of the baseline because given their work one *could* hold on to the baseline. Context sensitivity may well serve many masters.

In cases like (41) context supplies an ordering that tracks Inez's options and priorities; in cases like (42) context typically (but not invariably) supplies an ordering that tracks June's options and priorities. A natural challenge at this point is that this approach seems to stuff in a black box some systematic pattern that could be explained by the system. However, Bronfman and Dowell argue on the data side that the pattern is not *that* systematic after all

More recently, Daniel Skibra [Skibra, unpublished] has leveraged a number of observations at the interface of syntax and semantics into an innovative and linguistically informed approach to the Schroeder's challenge. Like the other critics of Schroeder, Skibra maintains that *ought* is not lexically ambiguous between a deliberative and an evaluative interpretation. Unlike those critics, however, he believes that there is a difference in logical form between deliberative *sentences* and *agentive sentences*. In this respect, Skibra's view resembles the AIP account. However, he deviates from AIP because he thinks that the availability of agent arguments depends crucially on whether the *embedded verb* can take an agent argument. If Skibra is right, progress on this problem will require much deeper integration between the study of deontic modality and the theory of argument structure in syntax and semantics.

7 Conclusion

I have little in the way of general conclusions at the end of this general survey. We have considered some important ways in which attention to natural language phenomena has driven intellectual progress in deontic logic. The topic surveyed here are by no means the only ones deserving on emphasis. Among omissions I will emphasize some last themes:

- free choice permission.[50]
- work on the performativity of *must*.[51]
- work on the "Britney Spears" problem for deontic conditionals.[52]
- work on anankastic conditionals.[53]

The area remains active and vibrant.

[50] [Kamp, 1973; Fusco, 2014; Starr, 2016; Willer, 2018].
[51] [Ninan, 2005; Mandelkern, forthcoming].
[52] [Zvolenszky, 2002; Zvolenszky, 2006; Cariani, 2013a; Carr, 2014].
[53] [Sæbø, 2001; Huitink, 2008; von Fintel and Iatridou, 2009; Condoravdi and Lauer, 2016].

References

[Aher, 2012] M. Aher. Free choice in deontic inquisitive semantics (dis). In *Logic, Language and Meaning*, pages 22–31. Springer, 2012.

[Arregui, 2010] A. Arregui. Detaching *if*-clauses from *should*. *Natural Language Semantics*, pages 241–293, 2010.

[Barker, 2010] C. Barker. Free choice permission as resource-sensitive reasoning. *Semantics and Pragmatics*, 3:10–1, 2010.

[Belnap *et al.*, 2001] N. Belnap, M. Perloff, and M. Xu. *Facing the Future*. Oxford University Press, 2001.

[Bledin, 2015] J. Bledin. Modus ponens defended. *The Journal of Philosophy*, 112(2):57–83, 2015.

[Bledin, 2020] J. Bledin. Fatalism and the logic of unconditionals. *Noûs*, 54(1), 2020.

[Bonevac, 1998] D. Bonevac. Against conditional obligation. *Noûs*, 32(1):37–53, 1998.

[Boylan, ms] D. Boylan. Putting 'ought's together. Texas Tech, ms.

[Brennan, 1993] V. M. Brennan. *Root and epistemic modal auxiliary verbs*. PhD thesis, University of Massachusetts, 1993.

[Bronfman and Dowell, 2016] A. Bronfman and J. Dowell. Contextualism about deontic modals. In Nate Charlow and Matthew Chrisman, editors, *Deontic Modality*, pages 117–142. Oxford University Press, 2016.

[Bronfman and Dowell, 2018] A. Bronfman and J. Dowell. The language of 'ought' and reasons. In Daniel Star, editor, *Oxford Handbook of Reasons and Normativity*. Oxford University Press, 2018.

[Cariani *et al.*, 2013] F. Cariani, M. Kaufmann, and S. Kaufmann. Deliberative modality under epistemic uncertainty. *Linguistics and Philosophy*, 36:225–259, 2013.

[Cariani, 2009] F. Cariani. *The Semantics of 'ought' and the Unity of Modal Discourse.* PhD thesis, UC Berkeley, 2009.

[Cariani, 2013a] F. Cariani. *Ought* and resolution semantics. Noûs, 47(3):534–558, 2013.

[Cariani, 2013b] F. Cariani. Epistemic and deontic *Should*. *Thought*, 2(1):73–84, 2013.

[Cariani, 2014] F. Cariani. Attitudes, deontics and semantic neutrality. *Pacific Philosophical Quarterly*, 2014.

[Cariani, 2016a] F. Cariani. Consequence and contrast in deontic semantics. *The Journal of Philosophy*, 113:396–416, 2016.

[Cariani, 2016b] F. Cariani. Deontic modals and probabilities: One theory to rule them all? In N. Charlow and M. Chrisman, editors, *Deontic Modals*, pages 11–46. Oxford University Press, 2016.

[Carmo and Jones, 2002] J. Carmo and AJI Jones. Deontic logic and contrary-to-duties. In *Handbook of philosophical logic*, pages 265–343. Springer, 2002.

[Carr, 2014] J. Carr. The *If P, Ought P* problem. *Pacific Philosophical Quarterly*, 95(4):555–583, 2014.

[Carr, 2015] J. Carr. Subjective ought. *Ergo*, 2(27):678–710, 2015.

[Charlow, 2013] N. Charlow. What we know and what to do. *Synthese*, 190:2291–2323, 2013.

[Charlow, 2016] N. Charlow. Decision theory: Yes! truth conditions: No! In N. Charlow and M. Chrisman, editors, *Deontic Modality*. Oxford University Press, 2016.

[Charlow, 2018] N. Charlow. Decision theoretic relativity in deontic modality. *Linguistics and Philosophy*, 41(3):251–287, 2018.

[Charlow, 2020] N. Charlow. Grading modal judgment. *Mind*, 129(515):769–807, 2020.

[Charlow, unpublished] N. Charlow. Modal triviality. unpublished.

[Chierchia *et al.*, 2008] G. Chierchia, D. Fox, and B. Spector. The grammatical view of scalar implicatures and the relationship between semantics and pragmatics. Harvard/MIT, 2008.

[Chisholm, 1963] R. Chisholm. Contrary-to-duty imperatives and deontic logic. *Analysis*, 24:33–36, 1963.

[Chrisman, 2012] M. Chrisman. On the meaning of 'ought'. In R. Shafer-Landau, editor, *Oxford Studies in Metaethics, vol. 7*, page 304. Oxford University Press, 2012.

[Condoravdi and Lauer, 2016] C. Condoravdi and S. Lauer. Anankastic conditionals are just conditionals. *Semantics & Pragmatics*, 9, 2016.

[Copley, 2006] Bridget Copley. What should *should* mean? unpublished manuscript, CNRS/University of Paris 8, 2006.

[Dowell, 2011] J. L. Dowell. A flexible contextualist account of epistemic modals. *Philosophers' Imprint*, 11:1–25, 2011.

[Dowell, 2013] J. L. Dowell. Flexible contextualism about deontic modals. *Inquiry*, 56(2-3):149–178, 2013.

[Feldman, 1986] F. Feldman. *Doing the Best We Can: An Essay in Informal Deontic Logic*. Dordrecht Reidel, 1986.

[Fine, unpublished] K. Fine. Chisholm's paradox and unconditional obligation. New York University, unpublished.

[Finlay and Snedegar, 2014] S. Finlay and J. Snedegar. One ought too many. *Philosophy and Phenomenological Research*, 84(1):102–124, 2014.

[Finlay, 2014] S. Finlay. *Confusion of Tongues*. Oxford University Press, 2014.

[Finlay, 2016] S. Finlay. Ought: Out of order. In N. Charlow and M. Chrisman, editors, *Deontic Modals*, pages 169–199. Oxford University Press, 2016.

[von Fintel and Iatridou, 2008] K. von Fintel and S. Iatridou. How to say *Ought* in foreign: the composition of weak necessity modals. In J. Guéron and J. Lecarme, editors, *Time and modality (Studies in Natural Language and Linguistic Theory 75)*, pages 115–141. Springer, 2008.

[von Fintel and Iatridou, 2009] K. von Fintel and S. Iatridou. What to do if you want to go to harlem? MIT, 2009.

[von Fintel, 1994] K. von Fintel. *Restrictions on Quantifier Domains.* PhD thesis, UMass, Amherst, 1994.

[von Fintel, 2012] K. von Fintel. The best we can (expect to) get? challenges to the classic semantics for deontic modals. presented at the 2012 Central APA, Chicago, IL, 2012. `http://mit.edu/fintel/fintel-2012-apa-ought.pdf`.

[Föllesdal and Hilpinen, 1971] D. Föllesdal and R. Hilpinen. Deontic logic: an introduction. In R. Hilpinen, editor, *Deontic Logic: Introductory and Systematic Readings*, pages 1–35. Reidel, 1971.

[Forrester, 1984] J. Forrester. Gentle murder,or the adverbial samaritan. *The Journal of Philosophy*, LXXXI(4):193–197, 1984.

[Fusco, 2014] M. Fusco. Factoring disjunctions out of deontic modal puzzles. In Fabrizio Cariani, Davide Grossi, Joke Mehens, and Xavier Parent, editors, *Deontic Logic and Normative Systems: DEON 2014*, pages 95–107. Springer Verlag, Berlin, 2014.

[Fusco, 2015] M. Fusco. Deontic modality and the semantics of choice. *Philosopher's Imprint*, pages 1–27, 2015.

[Gabbay, 2012] D. Gabbay. Temporal deontic logic for the generalised chisholm set of contrary to duty obligations. In *International Conference on Deontic Logic in Computer Science*, pages 91–107. Springer, 2012.

[Geurts, 2004] B. Geurts. On an ambiguity in quantified conditionals. unpublished, University of Njemegen, 2004.

[Goble, 1993] L. Goble. The logic of obligation, 'better' and 'worse'. *Philosophical Studies*, 70(2):133–163, 1993.

[Goble, 1996] L. Goble. Utilitarian deontic logic. *Philosophical Studies*, 82:317–357, 1996.

[Goble, 2013] L. Goble. Prima facie norms, normative conflicts, and dilemmas. In D. Gabbay, J. F. Horty, X. Parent, R. van der Meyden, and L. van der Torre, editors, *Handbook of Deontic Logic and Normative Systems. Vol. 1*, pages 241–351. College Publications, 2013.

[Hacquard, 2006] V. Hacquard. *Aspects of modality.* PhD thesis, MIT PhD, 2006.

[Hacquard, 2010] V. Hacquard. On the event-relativity of modal auxiliaries. *Natural Language Semantics*, 18(1):79–114, 2010.

[Hacquard, 2013] V. Hacquard. The grammatical category of modality. In *Proceedings of the 19th Amsterdam Colloquium.* 2013.

[Hilpinen and McNamara, 2013] R. Hilpinen and P. McNamara. Deontic logic: a historical survey and introduction. In D. Gabbay, J. F. Horty, X. Parent, R. van der Meyden, and L. van der Torre, editors, *Handbook of Deontic Logic and Normative Systems*, pages 72–86. College Publications, 2013.

[Holliday and Icard, 2017] W. Holliday and T. F. Icard. Indicative conditionals

and dynamic epistemic logic. In *Proceedings of TARK 2017*, pages 337–351, 2017.

[Holliday and Icard, 2018] W. Holliday and T. F. Icard. Axiomatization in the meaning sciences. In Derek Ball and Brian Rabern, editors, *The Science of Meaning*, pages 73–97. 2018.

[Horty and Belnap, 1995] J. F. Horty and N. Belnap. The deliberative *stit*: a study in the logic of action, omission, ability and obligation. *Journal of Philosophical Logic*, 24:583–644, 1995.

[Horty, 2001] J. F. Horty. *Agency and Deontic Logic*. Oxford University Press, 2001.

[Huitink, 2008] J. Huitink. *Modals, conditionals and compositionality*. PhD thesis, Radboud Universiteit Nijmegen, 2008.

[Jackson and Pargetter, 1986] F. Jackson and R. Pargetter. Oughts, options and actualism. *The Philosophical Review*, 95(2):233–255, 1986.

[Jackson, 1985] F. Jackson. On the semantics and logic of obligation. *Mind*, 94(374):177–195, 1985.

[Jackson, 1991] F. Jackson. Decision theoretic consequentialism and the nearest dearest objection. *Ethics*, 101(3):461–482, 1991.

[Jerzak, 2019] E. Jerzak. Two ways to want. *Journal of Philosophy*, 116(2):65–98, 2019.

[Jones and Pörn, 1985] AJI Jones and I. Pörn. Ideality, sub-ideality and deontic logic. *Synthese*, 65(2):275–290, 1985.

[Kamp, 1973] H. Kamp. Free choice permission. *Proceedings of the Aristotelian Society*, 74:57–74, 1973.

[Kaufmann, 2017] S. Kaufmann. The limit assumption. *Semantics and Pragmatics*, 10(18):1–27, 2017.

[Kolodny and MacFarlane, 2010] N. Kolodny and J. MacFarlane. Ifs and oughts. *Journal of Philosophy*, 107(3):115–143, 2010.

[Kratzer, 1977] A. Kratzer. What 'must' and 'can' must and can mean. *Linguistics and Philosophy*, 1(3):337–355, 1977.

[Kratzer, 1981] A. Kratzer. The notional category of modality. In B. Partee and P. Portner, editors, *Formal Semantics: the Essential Readings*. Blackwell, 1981.

[Kratzer, 1991a] A. Kratzer. Conditionals. In A. von Stechow & D. Wunderlich, editor, *Semantics: An International Handbook of Contemporary Research*. De Gruyter, 1991. from the *Semantics Archive*.

[Kratzer, 1991b] A. Kratzer. Modality. In A. von Stechow & D. Wunderlich, editor, *Semantics: An International Handbook of Contemporary Research*. De Gruyter, 1991.

[Kratzer, 2012] A. Kratzer. *Modals and Conditionals*. Oxford University Press, 2012.

[Lassiter, 2011] D. Lassiter. *Measurement and Modality: the Scalar Basis of Modal Semantics*. PhD thesis, NYU, 2011.

[Lassiter, 2016] D. Lassiter. Linguistic and philosophical considerations on bayesian semantics. In Nate Charlow and Matthew Chrisman, editors, *Deontic Modality*, pages 82–116. Oxford University Press, 2016.

[Lassiter, 2017] D. Lassiter. *Graded Modality*. Oxford University Press, 2017.

[Levinson, 2003] D. Levinson. Probabilistic model-theoretic semantics for want. In *Proceedings of SALT*, volume 22, pages 222–39. 2003.

[Lewis, 1974] D. Lewis. Semantic analyses for dyadic deontic logic. In *Papers in Ethics and Social Philosophy*. Cambridge University Press, 1974.

[Loewer and Belzer, 1983] B. Loewer and M. Belzer. Dyadic deontic detachment. *Synthese*, 54(2):295–318, 1983.

[MacFarlane, 2014] J. MacFarlane. *Assessment Sensitivity*. Oxford University Press, 2014.

[Mandelkern, forthcoming] M. Mandelkern. Practical Moore sentences. *Noûs*, forthcoming.

[McNamara, 2010] P. McNamara. Deontic logic. In E. N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Fall 2010 edition, 2010.

[Nair, 2014] S. Nair. Consequences of reasoning with conflicting obligations. *Mind*, 123(491):753–790, 2014.

[Nair, 2016] S. Nair. Conflicting reasons, unconflicting "ought"s. *Philosophical Studies*, 173(3):629–663, 2016.

[Ninan, 2005] D. Ninan. Two puzzles about deontic necessity. In Gajewski J., Hacquard V., Nickel B., and Yalcin S., editors, *New Work on Modality*. MIT Working Papers in Linguistics, 2005.

[Parfit, unpublished] D. Parfit. What we together do. Manuscript, Oxford University, unpublished.

[Portner, 2009] P. Portner. *Modality*. Oxford University Press, 2009.

[Prakken and Sergot, 1996] H. Prakken and M. Sergot. Contrary-to-duty obligations. *Studia Logica*, 57(1):91–115, 1996.

[Prakken and Sergot, 1997] H. Prakken and M. Sergot. Dyadic deontic logic and contrary-to-duty obligations. In *Defeasible deontic logic*, pages 223–262. Springer, 1997.

[Radford, 2004] A. Radford. *Minimalist Syntax*. Cambridge Textbooks in Linguistics. Cambridge University Press, 2004.

[Regan, 1980] D. Regan. *Utilitarianism and Cooperation*. Oxford University Press, 1980.

[Ross, 1941] Alf Ross. Imperatives and logic. *Theoria*, 7:53–71, 1941.

[Rubinstein, 2012] A. Rubinstein. *Roots of modality*. PhD thesis, UMass, Amherst, 2012.

[Sæbø, 2001] KJ Sæbø. Necessary conditions in a natural language. *Audiatur vox sapientiae: A Festschrift for Arnim von Stechow*, pages 427–449, 2001.

[Saint Croix and Thomason, 2014] C. Saint Croix and R. H. Thomason. Chisholm's paradox and conditional oughts. In F. Cariani, D. Grossi, J. Mehens, and X. Parent, editors, *Deontic Logic and Normative Systems:*

DEON 2014, pages 192–207. Springer Verlag, Berlin, 2014.

[Schroeder, 2010] M. Schroeder. Oughts, agents and actions. *The Philosophical Review*, 120(1):1–41, 2010.

[Silk, 2014] A. Silk. Evidence sensitivity in deontic modals. *Journal of Philosophical Logic*, 43:691–723, 2014.

[Silk, 2017] A. Silk. Modality, weights, and inconsistent premise sets. *Journal of Semantics*, 34(4):683–707, 2017.

[Silk, forthcoming] A. Silk. Weak and strong necessity modals. In B. Dunaway and D. Plunkett, editors, *Meaning, Decision, and Norms: Themes from the Work of Allan Gibbard*. Maize Books, forthcoming.

[Skibra, unpublished] D. Skibra. Ought and agency. Northwestern University, unpublished.

[Sloman, 1970] A. Sloman. Ought and better. *Mind*, 79(315):385–394, 1970.

[Snedegar, 2017] J. Snedegar. *Contrastive Reasons*. Oxford University Press, 2017.

[Stalnaker, 1975] R. Stalnaker. Indicative conditionals. *Philosophia*, 5:269–86, 1975.

[Starr, 2016] W. Starr. Expressing permission. In *Semantics and Linguistic Theory*, volume 26, pages 325–349, 2016.

[Szabó, 2015] Z. Gendler Szabó. Major parts of speech. *Erkenntnis*, 80(1):3–29, 2015.

[Thomason, 1981a] R. H. Thomason. Deontic logic and the role of freedom in moral deliberation. In R. Hilpinen, editor, *New Studies in Deontic Logic*, pages 177–186. Reidel, 1981.

[Thomason, 1981b] R. H. Thomason. Deontic logic as founded on tense logic. In R. Hilpinen, editor, *New Studies in Deontic Logic*, pages 165–176. Reidel, 1981.

[Van De Putte, 2018] F. Van De Putte. Coarse deontic logic. *Journal of Logic and Computation*, pages 1–32, 2018.

[Veltman, 1996] F. Veltman. Defaults in update semantics. *Journal of Philosophical Logic*, 25(3):221–261, 1996.

[Wedgwood, 2006] R. Wedgwood. The meaning of 'ought'. *Oxford Studies in Metaethics*, 1:127–160, 2006.

[Wedgwood, 2016] R. Wedgwood. Subjective and objective ought. In N. Charlow and M. Chrisman, editors, *Deontic Modals*, pages 143–168. Oxford University Press, 2016.

[Willer, 2012] M. Willer. A note on *Iffy* oughts. *Journal of Philosophy*, 109:449–461, 2012.

[Willer, 2014] M. Willer. Dynamic thoughts on ifs and oughts. *Philosopher's Imprint*, pages 1–27, 2014.

[Willer, 2016] M. Willer. Dynamic foundations for deontic logic. In N. Charlow and M. Chrisman, editors, *Deontic Modals*, pages 324–354. Oxford University Press, 2016.

[Willer, 2018] M. Willer. Simplifying with free choice. *Topoi*, 37(3):379–392, 2018.

[Yalcin, 2010] S. Yalcin. Probability operators. *Philosophy Compass*, pages 916–937, 2010.

[Yalcin, 2011] S. Yalcin. Nonfactualism about epistemic modality. In Egan A. and Weatherson B., editors, *Epistemic modality*. Oxford University Press, 2011.

[Yalcin, 2012] S. Yalcin. A counterexample to *Modus Tollens*. *Journal of Philosophical Logic*, 41:1001–1024, 2012.

[Zvolenszky, 2002] Z. Zvolenszky. Is a possible worlds semantics of modality possible? In B. Jackson, editor, *Proceedings of SALT 12*, pages 339–352. 2002.

[Zvolenszky, 2006] Z. Zvolenszky. A semantic constraint on the logic of modal conditionals. In B. Gyuris, L. Kálmán, C. Pi non, and K. Varasdi, editors, *Proceedings of the Ninth Symposium on Logic and Language (LoLa 9)*. Révai Digital Press, 2006.

Fabrizio Cariani
Philosophy Department
University of Maryland College Park

8
Deontic Logic and Ethics

Shyam Nair

Abstract. Though there have been productive interactions between moral philosophers and deontic logicians, there has also been a tradition of neglecting the insights that the fields can offer one another. The most sustained interactions between moral philosophers and deontic logicians have not been systematic but instead have been scattered across a number of distinct and often unrelated topics. This chapter primarily focuses on three topics. First, we discuss the "actualism/possibilism" debate which, very roughly, concerns the relevance of what one will do at some future time to what one ought to do at present (§2). This topic is also used to introduce various modal deontic logics. Second we discuss the particularism debate which, very roughly, concerns whether there can be any systematic general theory of what we ought to do (§3). This topic is also used to introduce various non-modal deontic logics. Third, we discuss collective action problems which concern the connection between the obligations of individuals and the behavior and obligations of groups of individuals (§4). This topic is also used to discuss formal systems that allow us to study the relationship between individuals and groups. The chapter also contains a general discussion of the relation between ethical theory and deontic logic

1 Deontic Logic, Ethical Theory, and Neutrality 552

2 Actualism and Possibilism 557
2.1 Modal Theories 560
2.1.1 SDL 560
2.1.2 Preference Semantics 562
2.1.3 Semantics with Agency 573

Thanks to the DEON community and especially Lou Goble, Jeff Horty, and Mark Schroeder who taught me almost everything I know about these topics. Of course, I am solely responsible for my remaining ignorance. Thanks also to Michael Gifford and Zach Horne for help with getting through the basic labor needed for a survey article of this sort. And finally thanks to the editors and especially Xavier Parent for their patience with me and to Jane Spurr for help with some of the figures.

Handbook of Deontic Logic and Normative Systems, Volume 2

2.2 An Insight From Ethics 579
2.3 Further Issues 583
2.3.1 Decision Rules in Modal Semantics 584
2.3.2 Other Interpretations 590

3 Particularism **592**
3.1 Motivations 593
3.2 The Formal Theory 597
3.3 Limitations of the Response to Particularism 606
3.4 Other Problems and Competing Implementations 609
3.4.1 Derivative and Non-Derivative Reasons 609
3.4.2 Undercutting Defeat and Downward Closure . . . 612
3.4.3 The Accrual of Reasons 616
3.4.4 Other Important Theories 618

4 Individual and Group Obligations **621**
4.1 Quantified Deontic Logic and Group Agency 623
4.1.1 Quantified Deontic Logic 624
4.1.2 Group Agency 627
4.2 A Speculative Alternative Framework 632
4.3 The Ethics of Existence 634

5 Further Topics **638**
5.1 The Logical Form of Obligation 638
5.2 Input/Output Logic 640
5.3 Contrary-to-Duty Obligation 643
5.4 Moral Conflicts 644

6 Conclusion **645**

Introduction

Though there have been productive interactions between moral philosophers and deontic logicians, there has also been a tradition of neglecting the insights that the fields can offer one another. The most sustained interactions between moral philosophers and deontic logicians have not been systematic but instead have been scattered across a number of distinct and often unrelated topics. This article will to some extent follow this approach.

After discussing a preliminary issue concerning the interaction between deontic logic and ethics (§1), the article focuses on three topics from ethics that are useful entry points for thinking about the interaction between ethics and deontic logic.[1] First, we discuss the "actualism/possibilism" debate which, very roughly, concerns the relevance of what one will do at some future time to what one ought to do at present (§2). Second we discuss the particularism debate which, very roughly, concerns whether there can be any systematic general theory of what we ought to do (§3). Third, we discuss collective action problems which concern the connection between the obligations of individuals and the behavior and obligations of groups of individuals (§4).

In discussing each of these issues, we begin with a philosophical discussion informed primarily by work in ethical theory. We then introduce a family of formal systems that offers an interesting perspective on these issues and consider how this is related to work in ethics. Each example introduces a different family of formal systems: the first introduces various modal theories; the second various non-modal theories; the third theories that allow us to study the relationship between individuals and groups. We by no means offer a systematic survey of these families of theories; instead, select representative examples are chosen and the reader is invited to explore these families in more detail by consulting other chapters of this handbook. The discussion of these sections closes by considering general features of these theories that are of philosophical interest even outside of our guiding topics. The article closes by discussing a grab bag of topics that are worthy of further consideration (§5)

But before we begin, four preliminary issues must be mentioned. First, my discussion of issues in ethical theory presupposes a certain limited conception of what ethical theory is. What I have in mind are topics of normative ethics, metaethics, and practical reason as they are typically discussed in contemporary analytic philosophy. This limited conception leaves out many other approaches that have legitimate claim to be called "ethical theory". But since I am most knowledgeable about these areas of philosophy, I focus on them in what follows. This is not intended to denigrate other approaches. And, to some extent, other chapters in this handbook explore some of these other approaches.

Second, I will not be discussing a number of issues in deontic logic and ethics that are of great interest and have received considerable

[1] These topics reflect my conception of the state of play rather than any disciplinary consensus on which topics are most important.

study. This is because this handbook has already dedicated whole chapters to these issues. They include discussion of moral conflicts [Goble, 2013]; right, duties, and other normative positions [Sergot, 2013]; and supererogation (McNamara, Chapter 3 in this Volume). I recommend to the reader each of these chapters and, for the most part, avoid discussing these issues. I also only briefly discuss contrary-to-duty obligations in this chapter (§5.3). This is because there was originally supposed to be a chapter dedicated to this topic in this volume. Unfortunately, the chapter did not come to fruition. My hope is that there will eventually be a handbook chapter written about this important topic.

Third, this article follows a traditional, though somewhat unfortunate, practice of treating "ought" and "should" as a term for discussing obligations. This does not, in my view, correspond to the ordinary English meaning of "ought" and "should". It is easiest to discuss obligation in English by using terms like "must" and "have to". Terms like "ought" and "should" (in their deontic uses) are most naturally used to discuss things that are optimal in some way. There are a number of issues in ethics where this distinction matters (e.g., the discussion of supererogation). But the topics that we pursue in what follows will not require this distinction so will tolerate our somewhat unfortunate way of speaking.

Fourth, this article is designed to be readable in a variety of ways other than straight through. Each section is self-contained (save for a few places and in these places, the reader is directed to the relevant background). In §2–4 after the introductory philosophical discussion, the next subsection on formal theories has subgroupings that may be read separately. Each subsection of §5 can be read on its own. For those uninterested in the core ethical cases, the subsection on the formal theories and the subsection of further discussion offers a general survey of other ethically relevant features of these theories. Those only interested in the core ethical cases may skip the subsection that describes further details of comparison and generalization of the formal theories. I recommend that the reader feel free to sample various subsections to see whether they are important for their purposes.

1 Deontic Logic, Ethical Theory, and Neutrality

We being with a general question: what is the relationship between deontic logic and ethics? We may make this question marginally but helpfully more precise by considering two questions:

(i) what kinds of intellectual interactions (borrowing of ideas, citations, collaboration, etc.) have occurred between individuals who studied deontic logic and individuals who have studied ethical theory?

(ii) what kinds of interesting relations (epistemological, metaphysical, conceptual, logical etc.) are there between the theories developed in these fields?

(i) is a question about intellectual history that is perhaps best studied by a detailed examination of the historical record. This is not the venue for such a study of the historical record. (ii), on the other hand, is a philosophical and formal question that can be studied using the tools that philosophers and logicians are familiar with. I will, therefore, primarily focus on this question.

One popular answer to the question has been that there is no particularly interesting relation between the theories of ethics and the theories of deontic logic. To the extent a theory of either field is promising, it neither tells in favor or against any theory in the other field. The dominant motivation for this view is a kind of neutrality thesis. To first approximation, the idea is that an adequate deontic logic should place no constraints whatsoever on what the correct ethical theory is. Indeed, it is a strike against a deontic logic if it is incompatible with a given ethical theory: the semantics of the logic should be compatible with various competing ethical theories; the set of valid arguments and theorems should be compatible with competing ethical theories. For example, if a proposed deontic logic somehow ruled out the possibility that it is wrong to lie to the murderer at the door, this logic should be rejected because it is incompatible with certain Kantian moral theories.[2]

There are different ways of motivating and developing this neutrality thesis. But I wish to begin with an analogy concerning classical predicate logic that will help to give us a flavor of what motivates the neutrality thesis. A familiar albeit disputable picture is that first-order predicate logic is a topic neutral system for understanding the commitments of any scientific or mathematical theory.[3] Logic should not decide between competing theories; scientific and mathematical methods should.

This perhaps initially innocuous sounding idea has substantial consequences. A simple example that illustrates this is that '$\exists x\ x = x$' is a theorem of classical first-order predicate logic. But some believe that

[2]See [Sayre-McCord, 1986] for an early discussion of this issue.

[3]But see [MacFarlane, 2000] for discussion of the difficulty of making good on this conception.

the neutrality constraint shows that this should not be a theorem of an adequate logic: It is a empirical matter whether there is at least one thing. Scientific methods not logic should settle this empirical matter.[4]

Analogously, the neutrality thesis that we are considering suggests that logic should not decide between competing ethical theories; moral argument should. So-called Standard Deontic Logic, for example, has it that '$O(a \wedge b) \rightarrow O(a)$' is a theorem.[5] But certain ethical theories reject this claim (as we will see in greater detail in §2.1). So, the argument goes, Standard Deontic Logic should be rejected. Logic should not decide on this matter; moral argument should.

Here it is important to see that the neutrality objection to Standard Deontic Logic is different from the so-called "paradoxes" that are used to object to Standard Deontic Logic. According to the neutrality objection, it does not matter whether the ethical theories that reject the theorem of standard deontic logic are true. All that matters is that they are genuine ethical theories (or perhaps that they are ethical theories that are not totally crazy). Returning to our analogy, the objector to standard first-order predicate logic of course concedes that '$\exists x\ x = x$' is true. What they claim is that since it is an empirical matter whether it is true, '$\exists x\ x = x$' is not a theorem. Similarly, even those who think '$O(a \wedge b) \rightarrow O(a)$' is true, the objector claims, should think '$O(a \wedge b) \rightarrow O(a)$' is not a theorem. Therefore, it does not actually matter whether we accept or reject the ethical theories that are incompatible with the theorem.

This, I hope, gives the flavor of the neutrality based argument for skepticism about the idea that there is an interesting relationship between deontic logic and ethical theory. This neutrality argument has been developed in more detail in a variety of ways. Each of these way of developing it involves spelling out a certain conception of what deontic logic is intended to capture. For example, according to some theorists, the truths of logic are a true in virtue of the meaning of the logical connectives alone. And, according to these theories, semantically competent users of these terms at some level know the meaning of these terms such that they are disposed to accept or at least not reject these truths (at least when they are clear-headed, thoughtful, not misspeak-

[4]Here I have in mind thoughts that motivate what [Nolt, 2018] calls "inclusive" or "universal free" logic.

[5]Though here we use italicized lowercase letters early in the alphabet for sentences and italicized lowercase letter late in the alphabet for variables, our conventions about these matters change at times. In particular, when discussing a formal system due to a particular author, I depart from these conventions and adopt the author's own notational style. Unfortunately, due to the diversity of approaches discussed in this article it is difficult to adopt a single uniform style throughout.

ing, etc.).[6] So, the argument goes, since there are semantically competent, clear-headed, thoughtful, intelligent, well-spoken moral philosophers who think that '$O(a \wedge b) \rightarrow O(a)$' is false and indeed even given arguments that it is false, '$O(a \wedge b) \rightarrow O(a)$' must not be a theorem.

Similar ideas have been developed by focusing on conceptual rather than semantic competence, by focusing on a special notion of logical competence, and by focusing on "metaethical" neutrality. Needless to say, these are deep philosophical waters and navigating them would require us to tackle fundamental issues concerning the relationship between logic, semantics, concepts, the a priori, and the necessary. Here is not the place to decide these issues.

Instead, I wish to just sketch two ways of responding to this concern. The first ways is conciliatory, relatively uncontroversial, and perhaps all that is needed for the purposes of continuing to take interest in the interaction between moral philosophers and deontic logicians. The second way is controversial and spells out an alternative conception of the role of deontic logic. While one need not adopt it to take interest in the issues discussed in the rest of the article, I include it because it is roughly the view that I accept and this view may have important consequences that are worth further consideration

The conciliatory view concedes to the objector their preferred conception of the role of logic. She distinguishes this notion of logic from another wider notion such as the one noted by John Burgess

> Among the more technically oriented 'logic' no longer means a theory about which forms of argument are valid, but rather means any formalism, regardless of intended application, that resembles a logic in the original sense enough to allow it to be usefully studied by similar methods. [Burgess, 2009, viii]

Or perhaps she only reserves the word 'logic' for the objectors intended conception and notes that one might nonetheless develop a formal system using tools that are similar to the ones used to develop a logic.

The conciliatory view, then, is that various proposed "logics" are really best interpreted (or at least may be interpreted) as formal systems that not properly called logics. So understood, they are not subject to the neutrality constraints. But of course, this also means they do not deserve the privileged status that real logic deserves according to the

[6] See [Carr, 2015; Cariani, 2016b; Charlow, 2016] for recent discussions of arguments like this one as well as several new arguments related to the linguistic semantics of 'ought' and the decision rules that they enforce (cf. §2.3.1)

objector. So these formal systems must contend with the arguments moral philosophers give.

In saying that these systems must contend with the arguments given by moral philosophers, we need not concede that the formal properties of the system are not themselves of interest. These properties may well speak to the explanatory power, simplicity, and coherence of the resulting theory and thereby be evidence for the theory that must be considered together with the specifically moral arguments given by ethical theorists. The resulting picture, then, is one on which ethical theory and deontic logic understood as formal system building are relevant to one another though neither enjoys privilege. They are engaged in the same, or a least substantially similar, project even though they typically emphasize different topics and employ different tools.

The idea that theories in deontic logic can be interpreted this way and that so-interpreted there is an interesting relationship between deontic logic and ethical theory should, I think, be uncontroversial. And it is all that is needed to see the interest in the various topics that we will discuss below.

Nonetheless, some may believe the deontic logic can and should also be interpreted in the stricter sense and so-understood it has important contributions to make. What's more, logic in the strict sense has a kind of privileged status. There are special data that is intended to capture and it has special privilege over certain other domains. This is a view on which deontic logic has priority over ethical theory.

I do not accept this "logical priority" view and will not defend it here. But — and this is my second point about this neutrality based objection — I do believe that deontic logic can and often should be interpreted as being logic in whatever strict sense there is of that term and even so understood the neutrality objection fails. The mistake, I believe, is interpreting logic in the strict sense to somehow capture a set of truths no competent clear headed person would reject. I instead reject the idea there is any such set of truths to be had.[7] Logic is simply a very abstract theory of reality. As such it is responsible to the totality of evidence including arguments from moral philosophy (as well as semantics, the sciences, etc.). But similarly, moral theories also are beholden to the evidence and argument for a given deontic logic. And, as I mentioned earlier, often the formal features of theories allow us to say quite a lot about the explanatory power, simplicity, and coherence

[7]The conception of logic that I accept here is a form of "anti-exceptionalism about deontic logic" akin to the view of philosophical methodology given in [Williamson, 2008] and discussed as it applies to modal logic in [Williamson, 2015, 423–429].

of a theory.

This response, then, is a wholesale rejection of the neutrality thesis. Its relevance here is that if it is right (and I don't pretend to have given an argument that it is right), the search for some neutral real logic is simply misguided. This consequence if correct undercuts the motivation for a number of logics that have been developed or at least requires that these motivations be understood differently.

In any case, the approach taken in this article will be one on which neither ethics nor deontic logic has any kind of priority. Both are beholden to the same body of evidence and aim to understand a common subject matter even if their approaches are different.

2 Actualism and Possibilism

We now turn to considering three debates in moral philosophy that are especially closely related to issues in deontic logic. The debate that we will look at in this section is the one between so-called "actualism" and "possibilism".[8]

It is easiest to introduce the issue by considering an example such as the following one due to Michael Zimmerman:

> I have been invited to attend a wedding. The bride-to-be is a former girlfriend of mine; it was she who did the dumping. Everyone, including me in my better moments, recognizes that she was quite right to end our relationship; we were not well suited for one another, and the prospects were bleak. Her present situation is very different; she and her fiancé sparkle in one another's company, spreading joy wherever they go. This irks me to no end, and I tend to behave badly whenever I see them together. I ought not to misbehave, of

[8]The classic discussions of these kinds of cases include [Feldman, 1986; Goldman, 1976; Goldman, 1978; Greenspan, 1978; Jackson, 1985; Jackson and Pargetter, 1986; Sobel, 1976]. More recent discussions include [Bykvist, 2002; Carlsson, 1999a; Carlsson, 1999b; Cohen and Timmerman, 2016; Curran, 1995; Gustafsson, 2014; Jackson, 2014; Portmore, 2011; Portmore, 2013; Portmore, 2017; Portmore, 2019; Ross, 2012; Timmerman, 2015; Timmerman and Cohen, 2016; Vessel, 2003; Vessel, 2009; Vessel, 2016; Zimmerman, 1996], and [Zimmerman, 2006]. These examples (together with other data) have also been used to motivate some interesting proposal in natural language semantics such as [Cariani, 2013; Snedegar, 2014], and [Snedegar, 2017]. [Cariani, 2016a] is especially worth consulting as it shows there are a variety of ways these proposal in natural language semantics can define validity and each choice has its unique set of costs. See Chapter 7 in this Volume for more discussion of proposal in natural language semantics.

> course, and I know this; I could easily do otherwise, but I do not. The wedding will be an opportunity for me to put this boorishness behind me, to grow up and move on. The best thing for me to do would be to accept the invitation, show up on the day in question, and behave myself. The worst thing would be to show up and misbehave; better would be to decline the invitation and not show up at all. [Zimmerman, 2006, 153]

In this case, should I accept the invitation or not? So far the answer may look straightforward. But Zimmerman adds an important wrinkle to the case: suppose that "if I accepted the invitation, I would show up and misbehave (whereas I would not do this if I declined). I need not misbehave (for, as noted, I could easily do otherwise); nonetheless, this is what I would in fact do." (ibid. 153).

In this setting, an interesting case can be made for each answer to the question of whether to accept the invitation. Some — the possibilists — claim that I ought to accept the invitation. They point out that it is perfectly possible for me to accept the invitation and behave myself and this would lead to the best outcome. And possibilist think that in deciding whether to do an act we should consider the best outcome one *can* bring about that involves that act. So, according to the possibilist, I ought to accept.

Others — the actualists — claim that it is not the case that you ought to accept. They point out that while I, of course, can behave myself, I, in fact, won't behave myself if I were to accept the invitation. So if I were to accept, I would actually bring about the worst outcome. And actualist think that in deciding whether to do an act, we should consider the outcome that *would* result if one did the act. So, according to the actualists, it is not the case that I ought to accept.

This actualist verdict is a commitment of standard forms of consequentialism

> S ought to do x iff the outcome of S's doing x is better than the outcome of S's refraining from doing x

where the outcome of S's doing x is understood as follows:

> a possible world w is the outcome of S's doing x iff if S were to do x, then w would obtain[9]

[9]This definition makes most sense in a setting where the so-called uniqueness assumption holds. We discuss theories that embrace alternatives to this assumption below.

So understood, it is not the case that I ought to accept. This is because if I were to accept, the resulting possible world would be one in which I go to the wedding and misbehave. This outcome is worse than the outcome of refraining from accepting the invitation. If I were to refrain from accepting the invitation, the resulting possible world would be one in which I do not go to the wedding and so do not misbehave. Thus according to consequentialism it is not the case that I ought to accept.

What is at stake in this debate is not just some claims about what we ought to do in certain examples. These examples are, instead, intended to make vivid different perspectives on whether and how what one will do in the future can affect what one ought to do presently; different perspectives on how facts about one's future agency can affect what one ought to do presently.

Of course, no serious deontic logic "out of the box" has a commitment about whether one ought to accept or reject the invitation. Indeed, all serious deontic logics are compatible with the actualist claim that it is not the case that I ought to accept and the possibilist claim that I ought to accept. But there is another more subtle commitment of actualists and standard consequentialism that is at odds with certain deontic logics.

To see this, consider the conjunctive act of accepting the invitation and behaving myself at the wedding. Now this, by stipulation of the case, is something that I am able to do. Since this would bring about the best possible outcome, the possibilist say that I ought to accept the invitation and behave myself.

But what does the actualist and the standard consequentialist say about this conjunctive act? To see what they say, we need to consider what would result from doing the conjunctive act. We know that what would result is the best outcome. So according to actualism (and standard consequentialism), I ought to accept the invitation and behave myself.

Thus, actualists and standard consequentialists are committed, then, to the claim that I ought to accept the invitation and behave myself ($O(accept \wedge behave)$) and that it is not the case that I ought to accept the invitation ($\neg O(accept)$). This means actualists are committed to denying the following principle:

DEONTIC INHERITANCE: if p entails q, then Op entails Oq

In what follows, we will explore how this controversy looks from the perspective of deontic logic. We use this as an opportunity to discuss deontic logics that are modal logics as they are most closely connected

to the consequentialist way of thinking and to the cases that we are interested (§2.1). We then return to some work in ethics about the possibilism and actualism debate and note some ways this work may contribute to the work in deontic logic (§2.2). We close with a more general discussion of some features of these theories (§2.3).

2.1 Modal Theories

We begin (in §2.1.1) by presenting the so-called Standard Deontic Logic (SDL). We then consider theories in the preference deontic logic tradition (§2.1.2). And finally theories that incorporate special representations of agency as well as preferences (§2.1.3). As we will see, SDL is not an especially helpful model of our cases. But it paves the way for more sophisticated theories that tell us something more interesting.

2.1.1 SDL

Let us remind ourselves of the Kripke semantics for modal logic. A semantic structure is a triple $\langle W, R, V\rangle$ where W is understood to be a non-empty set of possible worlds, R is understood to be a relation on W and V is function from sentence-world pairs to truth values that obeys the usual rules for non-modal vocabulary and the following additional rule for the modal operator, $\Box$:

> $\Box p$ is true at w relative to $\langle W, R, V\rangle$ iff for all w' such that wRw', p is true at w'

Structures like this where we put no constraint on what properties relation R has provide a sound and complete semantics for the axiom system known as **K**. Axiom system **K** consists of the theorems of propositional logic, axiom K that say $\Box(p \to q) \to \Box p \to \Box q$, the rule of necessitation that says if $\top$ is a theorem, one may infer $\Box\top$, and the rule of modus ponens.

When investigating metaphysical possibility, certain additional axioms are interesting to consider such as this:

$$\Box p \to p$$

This axioms is holds in every structure where R is a reflexive relation in the sense that for all w, wRw. But this axiom is undesirable in a deontic context where we interpret '$\Box$' as 'it is obligatory that': it is an obvious albeit unfortunate fact that certain things that ought to obtain fail to obtain.

On the other hand, in the deontic setting an attractive idea is:

$$\Box p \rightarrow \Diamond p$$

where '$\Diamond$' is defined so that it is equivalent to '$\neg\Box\neg$'. In the deontic setting where we interpret '$\Box$' as 'it is obligatory that' and '$\Diamond$' as 'it is permitted that', this axiom expresses the natural thought that that which is obligatory is also permitted. If we add this axiom to the system **K**, and add the restriction that R is a serial relation in the sense that for all w, there is a w' such that wRw', the semantics is sound and complete for this system. This system is often called "Standard Deontic Logic" or SDL.

How may the formal structure of SDL be interpreted? For each world, SDL assigns it a non-empty set of wolds[10] and anything that is true in every world in this set is what ought to be. The key question then is how we can understand this set of worlds in a sensible way; what must this set of worlds be like for it to be the right kind of thing to witness the truth of various deontic claims.

One simple idea is that this set of worlds is somehow deontically ideal. Now in moral philosophy one does not typically encounter moral theories that determine what is obligatory by citing deontically ideal ways that the world could be. But moral theories can be understood as determining such a set. For example, a consequentialist theory can be understood as saying that the deontically ideal worlds are the worlds that contain the most value. And perhaps other theories too can be regimented in this way. Perhaps Kantians can understand the set of ideal worlds as the set of worlds where the categorical imperative is obeyed. Perhaps contractualist can understand it as the set of world in which the rules no one can reasonably reject are obeyed. Perhaps virtue theories can understand it as the set of the worlds where things are as they would be if we were fully virtuous.

Each of these interpretations has some initial plausibility and the structure of SDL does not obviously rule one out or rule another in. But there are, in fact, non-trivial constraints imposed by the SDL semantics.

We can see this by considering what a consequentialist interpretation of SDL might be. It is natural to think of the set of ideal worlds relative to some particular world as the best or most valuable available worlds. While this may be a sensible interpretation of SDL that is inspired by taking goodness to be important in much the way the consequentialist takes goodness to be important, the resulting theory is inconsistent with standard act consequentialism.

[10]Each world w is assigned $\{w' \mid wRw'\}$ which is non-empy because R is serial.

As we noted earlier, standard act consequentialism rejects DEONTIC INHERITANCE (which, recall, says that if p entails q, then Op entails Oq). SDL, on the other hand, embraces it. To see this, assume that p entails q and that Op holds. According to SDL, Op holds because each of the best worlds is a world in which p. Since p entails q, then in each of these worlds q is also true. Thus, according to SDL, Oq holds because each of the best worlds is a world in which q. So SDL is incompatible with standard act consequentialism and with actualism.

SDL however is compatible with the possibilist verdict. Should we conclude from this that possibilism is correct or that SDL is incorrect? I do not believe so. SDL offers so little guidance about what to think about this concrete case that it is hard to see it as a definitive argument against the actualist view. It is also hard to take seriously the idea that consequentialism, actualism, and their verdicts about Zimmerman's wedding case provide a decisive objection to SDL. After all, consequentialism and actualism are themselves extremely controversial views and it is a contested matter what to say about Zimmerman's case.

I think instead we can draw two morals. First, we need to continue to look for a logic that might tell us something more interesting about the actualism/possibilism debate and in particular we may wish to look for a logic that is compatible with actualism. Second, this example should teach us that to test a moral theory's compatibility with a given semantics it is not enough to find a first pass gloss on the semantics that fits with the theory. Each semantics imposes some structural constraints and these must be checked to see if they fit the theory.[11]

Let's turn now to richer logics that may give us some more insight into our target cases.

2.1.2 Preference Semantics

Perhaps the logic that is most closely connected with the consequentialist and actualist view is the one developed in and Lou Goble's "A Logic of *Good*, *Should*, and *Would* Part I" ([Goble, 1990a]) and "A Logic of *Good*, *Should*, and *Would* Part II" ([Goble, 1990b]). Here I rehearse a simplified version of Goble's framework. Later we will consider Hansson's logic that puts goodness or preference at center stage as well.[12]

[11] §2.3.2 discusses some limitations of non-value-based interpretations even if they pass this test.

[12] There are many other preference-based frameworks for deontic logic, but we focus on these two frameworks because they are perhaps the most directly relevant to our discussion. Two early discussions include [Jennings, 1974] and [Lewis, 1974].

Goble's theory, like consequentialism, appeals to both facts about counterfactuals and facts about the ordering of outcomes according to their goodness to determine what is obligatory. More precisely, for a given world w_i, we have a comparative similarity ordering on worlds, $\leq_{w_i}$, and a comparative betterness ordering on worlds $\mathbf{Bt}_{w_i}$. We read '$w_j \leq_{w_i} w_k$' as saying 'w_j is at least as similar to w_i as is w_k'. We assume that this ordering is transitive (i.e., if $w_j \leq_{w_i} w_k$ and $w_k \leq_{w_i} w_l$, then $w_j \leq_{w_i} w_l$) and connected (i.e., for any w_j, w_k either $w_j \leq_{w_i} w_k$ or $w_k \leq_{w_i} w_j$). We read '$w_j\mathbf{Bt}_{w_i}w_k$' as saying 'w_j is, from the perspective of w_i, strictly better than w_k'. We assume that this ordering is asymmetric (if $w_j\mathbf{Bt}_{w_i}w_k$, then it is not the case that $w_k\mathbf{Bt}_{w_i}w_j$) and transitive (if $w_j\mathbf{Bt}_{w_i}w_k$ and $w_k\mathbf{Bt}_{w_i}w_l$, then $w_j\mathbf{Bt}_{w_i}w_i$).

Two additional and more controversial assumptions are needed about the comparative similarity ordering. First, we make the "strong centering" assumption that each world is uniquely most similar to itself (more formally: for each world w_i, $w_i \leq_{w_i} w_j$ for every w_j and if there is a w_j such that $w_j \leq_{w_i} w_i$, then $w_j = w_i$). Next we will use $\leq_{w_i}$ to define the notion of the A-alternative to w_i, $[A]_{w_i}$,as follows:

> $[A]_{w_i} = \{w_j \mid A$ is true at w_j and for all w_k such that A is true at w_k, $w_j \leq_{w_i} w_k\}$

In the case where A is true at w_i, $[A]_{w_i} = w_i$ because of the strong centering assumption that we have made. In the case where A is not true at w_i, we make the "limit" assumption about the ordering which requires that $[A]_{w_i}$ is always a non-empty set of worlds.[13]

This allows us to define an obligation operator. Goble defines the operator to ensure that it only applies to contingent claims. I ignore the complexity added to the definition to ensure this but will discuss applications of Goble's theory involving only contingent claims below. So a simplified version of Goble's idea is this:

> $O(A)$ is true at w_i iff for all $w_j \in [A]_{w_i}$ and for all $w_k \in [\neg A]_{w_i}$, $w_j Bt_{w_i} w_k$

This is natural formalization of the consequentialist idea. The w_j's and the w_k's represent the outcomes of A and $\neg A$ respectively. An act is obligatory exactly if its outcomes are better than the outcomes of its negation. The notion of outcome is given to us through counterfactuals because $[A]_{w_i}$ is defined using the comparative similarity ordering that we use to evaluate counterfactuals.

[13]We discuss further issues related to these assumptions in §2.3.1

This theory, as we might suspect, does not validate DEONTIC INHERITANCE (which recall is that if P entails Q, then $O(P)$ entails $O(Q)$). More interestingly however, the theory allows us to explore the formal structure of cases that are more complex variants of our wedding case. As the wedding case is told, we know that there are two possibilities where one accepts. In one possibility, one behaves and the other one does not behave. We know the first is the best possibility. We also know there is a possibility where one does not accept. In the standard telling there is one such possibility (that is relevant). But we can consider other more complex cases where there are two possibilities where one does not accept. In one possibility, one behaves; in the other, one does not. Suppose that the rankings of these possibility is that the best possibility is one in which one accepts and behaves, the next is one in which one does not accept and behaves, the third is one in which one accepts and does not behave, and the worst is one in which one does not accept and does not behave. What ought to be done in this more complex case?

According to the actualist, the answer to this question depends on what one would do. We know that if I were to accept, I would misbehave. So it must be that I do not accept and behave at the actual world. This follows from our strong centerting assumption which says that if I in fact accept and behave, then if I were to accept, I would behave. This gives us one (perhaps obvious) constraint on the formal analysis of the example. The actual world is not one in which I both accept and behave. This leaves us with three potentially interesting different ways the actual might be:

$$@_1 : \neg A \wedge \neg B \qquad @_2 : A \wedge \neg B \qquad @_3 : \neg A \wedge B$$

where 'A' stands for 'I accept the invitation', where 'B' stands for 'I behave', and where the numerical subscript tells us how valuable each world is (where larger the number, the more valuable). We further know that the world in which I accept and behave is not "nearby" in the relevant sense. Table 1, then, summarizes different ways this case might play out.

The spatial distance from the actual world gives us the similarity ordering on worlds relative to the actual world. The numerical subscript tells you the value of each world so that the world whose subscript is a higher number is a more valuable world. So the most valuable world is the world where I accept and behave, followed by a world where I don't accept and behave, followed by a world in which I accept and don't behave, and finishing with a world in which I don't accept and don't behave. Notice the world in which I accept and behave is always the

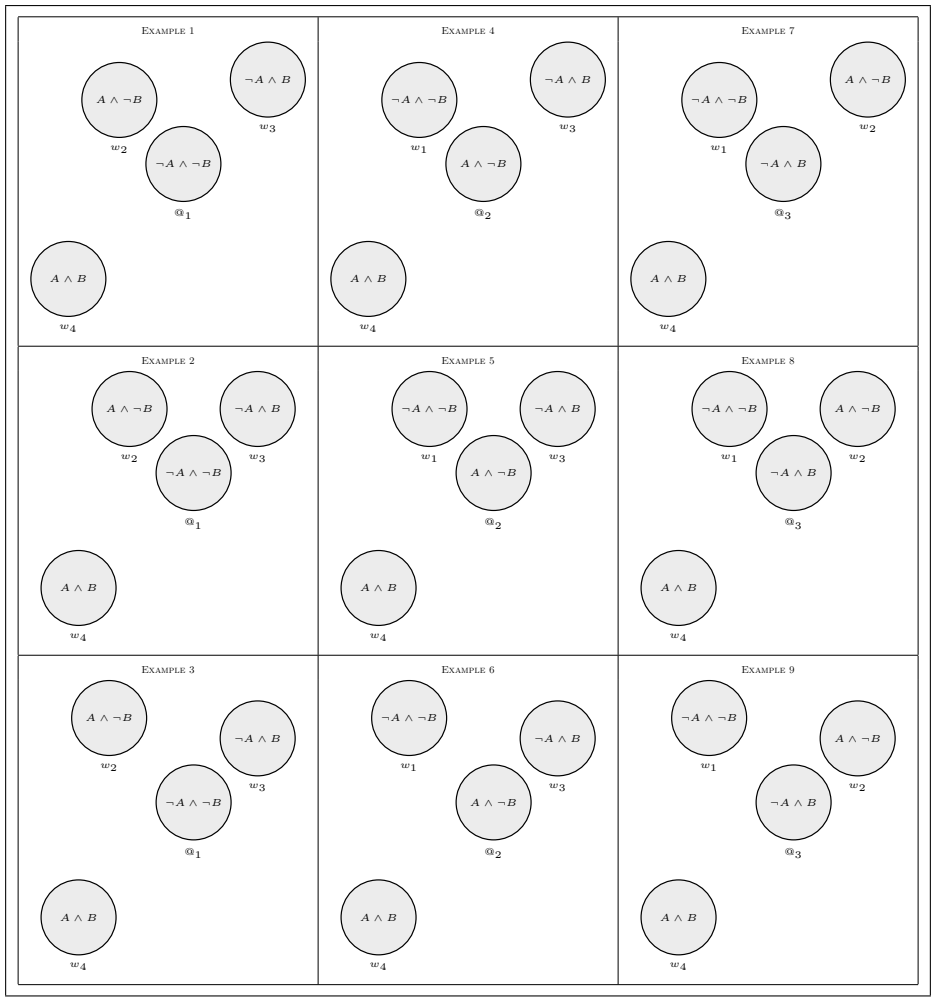


Figure 1: The dependence of what we ought to do on what we actually do in Goble's theory.

furthest from the actual world so the counterfactual "If I were to accept, then I would not behave" comes out true in all examples.

Looking at these example and checking definitions, we can see that in all of them I ought to accept and behave and in all them I ought to behave. But consider whether I ought to accept. In Example 1–4, I ought to accept. But in Example 5–9, it is not the case that I ought to accept. In Example 5–9, consider whether I ought to not accept or whether I am merely permitted to not accept. As it turns out, in Example 6–9 I ought to not accept. And only in Example 5 am I merely permitted to not accept.

These results are not, on reflection, especially surprising. But they are worth noticing. We already know that consequentialism leads to a failure of DEONTIC INHERITANCE because we know that $O(accept \wedge behave)$ but $\neg O(accept)$ (and Examples 5–9 also show this). But looking at Goble's natural formalization of consequentialism allows to see that the theory also rejects a number of other natural principles:

> NO CONFLICTS: $\neg(O(A) \wedge O(B))$ if $\{A, B\}$ is inconsistent
> NO PERMISSIBLE CONFLICTS: $\neg(O(A) \wedge P(B))$ if $\{A, B\}$ is inconsistent
> AGGLOMERATION: If $O(A)$ and $O(B)$, then $O(A \wedge B)$

To see this notice that in examples 6–9, we have that $O(A \wedge B)$ and $O(\neg A)$ even though $\{A \wedge B, \neg A\}$ is inconsistent so we have a counterexample to NO CONFLICTS. Next recall I mentioned as an aside the Goble's official definition of obligation blocks tautologies and contradictions from being obligatory. So we know that $\neg O((A \wedge B) \wedge \neg A)$. Thus, these examples are also counterexamples AGGLOMERATION. Finally we can notice that in example 5 we only get a counterexample to NO PERMISSIBLE CONFLICTS.

Nonetheless, Goble's theory is provably consistent and avoid trivializing the logic in various ways. It also validates certain restricted principles such as the following:

> NO STRICT CONFLICTS: $\neg(O(A) \wedge O(\neg A))$

And Goble's full theory has the resources to represent evaluative claims about what is good and bad and he proves a number of interesting theorems connecting these evaluative claims with claims about what we ought to do. Goble's logic, then, shows one way to have a coherent actualist and consequentialist theory.

Returning again to Table 1, notice that since the ranking of the outcomes according to how good they are is the same in all the cells, the differences between what is obligatory in each illustrates the dependency of what we ought to do on how similar worlds are to one another. Since according to strong centering, the nearest world is always the actual world, this in turn means what ought to be done partially depends on what is done. So for instance compare Example 2, 5, and 8. Example 2 and 8 feature a world in which we accept and a world in which we don't accept that are equally similar to the actual world. They feature distinct actual worlds. This results in the counterfactual 'If I were to not accept, then I would not behave' being true in Example 2 but false in 8. Similarly, the counterfactual 'if I were to not accept, then I would

behave' is true in Example 8, but false in 2. Example 5 is an intermediate case where both counterfactuals are false but 'if I were not to accept, then I might behave' and 'if I were not to accept, then I might not behave' are both true. These are of course plausible claims about which counterfactuals are true in these cases.

But this in turn percolates up to what is obligatory in Goble's theory. It makes it so that in Example 2 one ought to accept, in Example 5 one is merely permitted to accept and merely permitted to not accept, and in Example 8 one ought to not accept. And this dependence of obligation on what is actually done does not quite have the same initial plausibility. One, perhaps tendentious, way to put this concern is that in Example 2 and 8 one takes seriously the possibility in which one accepts and does not behave. But in both examples there is another possibility in which one does not accept (in Example 2, the possibility that one does not accept and behaves and in Example 8, the possibility that one does not accept and does not behave) that is equally close to the actual world that one ignores in evaluating what one ought to do. Instead, one merely focuses on what one actually does. There is something, at least initially, strange about this. We will see later that critics of actualism and defenders of something like DEONTIC INHERITANCE in moral philosophy have developed this suspicion that something is strange about this kind of dependence in detail. We will look at this in the §2.2 when I return to discussing some of the developments in ethics.

We now turn to a theory developed by Sven Ove Hansson 2001, 2013 that shares similarities with Goble's theory but, as we will see, is also interestingly different. We follow the presentation of Hansson's theory given in volume one of this handbook [Hansson, 2013]. Hansson's theory makes use of the idea of a preference relation on worlds. This is similar to Goble's idea of using a **Bt** ordering on worlds. But Hansson's view differs from Goble's in two respects. First, it does not make use of the notion of comparative similarity. Second, Hansson's view can be developed so that the ordering of worlds is a consequence of an ordering over propositions. And indeed, it is this development that is of most interest to us here.

So for Hansson we start with an ordering on propositions that is assumed to be transitive and complete. We read '$p \geq q$' as 'p is weakly preferred to q' and use $>$ for its strict counterpart (i.e., $p > q$ iff $p \geq q$ and $q \ngeq p$). Using this, Hansson is able to define a family of obligation operators. To do so, one selects a threshold proposition f and says $O(p)$ iff $f \geq \neg p$. So if $\neg p$ is at or below some threshold, then it is

obligatory that p. The threshold can be thought of as a "least" forbidden proposition. This is because it is forbidden that p,$F(p)$, is equivalent to $O(\neg p)$. Since $O(\neg p)$ iff $f \geq p$, $F(p)$ iff $f \geq p$. Thus any proposition q such that $q > f$, $\neg F(q)$. So in this sense, f is the least forbidden proposition.

So understood the semantics invalidate both DEONTIC INHERITANCE (which recall is that if p entails q, $O(p)$ entails $O(q)$) and AGGLOMERATION (which recall is if $O(p)$ and $O(q)$, then $O(p \wedge q)$). Indeed it validates no interesting theorems that can be stated solely in deontic vocabulary. But one may add structure of $\geq$ and study the results. One particularly interesting property that Hansson disusses is this:

$\geq$ is interplorative iff $(p \geq (p \vee q) \geq q)$ or $(q \geq (p \vee q) \geq p)$

This in turn is equivalent two other notable conditions:

$\geq$ is interpolative iff $\geq$ satisfies the following two conditions:
(a) $p \geq (p \vee q)$ or $q \geq (p \vee q)$ and (b) $(p \vee q) \geq p$ or $(p \vee q) \geq q)$

The idea of interplorativity says a disjunction can be as preferred as one of its disjuncts or as dispreferred as one of disjuncts or anything in between. What it prohibits is a disjunction being more prefered than both of its disjuncts or more disprefered than both of its disjuncts.

If we accept this assumption, the resulting logic has more structure. In particular if accept clause (a), we ensure that AGGLOMERATION holds. If we accept clause (b) we ensure the following principle holds:

DISJUNCTIVE DIVISION: if $O(a \wedge b)$, then $O(a)$ or $O(b)$

Nonetheless we still do not have DEONTIC INHERITANCE. To see this, consider the following ordering:

$$p > f > (p \vee q) \geq q$$

Obviously, this satisfies interpolativity. Additionally suppose that the ordering is such that logically equivalent formula occur in the same place. Now we have $O(\neg p \wedge \neg q)$ because $\neg(\neg p \wedge \neg q)$ is logically equivalent to $(p \vee q)$ and $f > (p \vee q)$. But we have $\neg O(\neg p)$ because $\neg\neg p$ is equivalent to p and $p > f$. Thus, even though $\neg p \wedge \neg q$ entails $\neg p$ and $O(\neg p \wedge \neg q)$, we do not have $O(\neg p)$.

How shall we interpret this preference relation? The name suggests one obvious interpretation is that the preferences are the preferences of the agent in question. Another natural interpretation that fits will with the consequentialist perspective is that it is a goodness ordering.

Whichever interpretation we might choose, however, we need to ask whether it vindicates interpolativity. There is some initial plausibility to the idea that the better-than relation does. Suppose p is the proposition that it is a sunny day and q is the proposition that it is a rainy day. In this case, it hard to see how the the proposition that it is a sunny day or a rainy day could be strictly worse or strictly better than both of these propositions. That said, the assumption is not trivial and is not satisfied by a variety of interpretation. For example, as [Hansson, 2013]: 490-1 observes, expected utility generates preference orderings that do not satisfy interpolativity.

In any case, Hansson's theory, like Goble's theory, is capable of vindicating the actualist claim that you ought to accept and write, but it is not the case that you ought to accept. To see this, consider this ordering:

$$\neg accept > f > (\neg accept \vee \neg behave) \geq \neg behave$$

Once again suppose that the ordering is such that logically equivalent formula occur in the same place. Now we have $O(accept \wedge behave)$ because $\neg(accept \wedge behave)$ is logically equivalent to $(\neg accept \vee \neg behave)$ and $f > (\neg accept \vee \neg behave)$. But we have $\neg O(accept)$ because $\neg accept > f$. Further, we have $O(behave)$ because $f > \neg behave$.

Thus, Hansson and Goble both can get the result that $O(accept \wedge behave)$ and $\neg O(accept)$. Interestingly however, Goble's theory, but not Hansson's, allows for there to be cases where $O(accept \wedge behave)$ and $\neg O(accept)$ and $\neg O(behave)$. Since Hansson's theory validates DISJUNCTIVE DIVISION from above (if $O(a \wedge b)$, then $O(a)$ or $O(b)$) it cannot allow for this. One might expect given our discussion earlier that this is because Goble does not require that **Bt** to be interpolative. But, in fact, Goble's theory predicts failures of DISJUNCTIVE DIVISION even in settings where the ordering satisfies interpolativity. Let's look at this in detail

To do this, we need to fill in the wedding case differently than we have so far. Figure 2 summarizes the situation.

As we can see, what I actually do is stay at home and behave. And just as in the original telling, if I were to accept, I would go to the wedding and misbehave and that would be the worst thing of all. And once again as the original telling goes, if the agent were to accept and behave, I would go to the wedding and everything would great so that is better than both of these outcomes.

But there is an important difference with this example. It is one where if I were to misbehave, I would in fact do it at home having

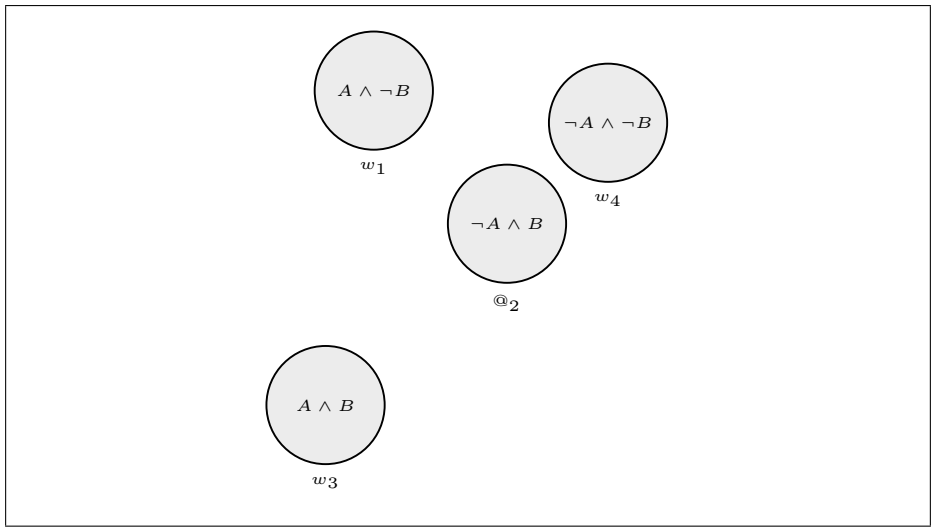


Figure 2: Interpolativity without Disjunctive Division in Goble's Theory

not accepted the invitation and not gone to the wedding. Finally, for whatever reason this is best outcome of all. This information will allow us to see that Goble's theory does not validate DISJUNCTIVE DIVISION.

According to Goble, in order to check whether we ought to accept and behave. We need to compare the nearest world or worlds in which you do accept and behave to the nearest world or worlds in which that conjunction is false; if those *accept* $\wedge$ *behave*-worlds are uniformly better than those where the conjunction is false, you ought to accept and behave. We can see that the nearest world and only world in which one accepts and behaves is w_3 and the nearest world in which this conjunction is false is the actual world, @$_2$. Finally, w_3 is better than @$_2$ so we have $O(accept \wedge behave)$.

Next consider whether you ought to accept. We can see that the nearest *accept*-world is w_1 and the nearest $\neg$*accept*-world is @$_2$. But @$_2$ is better than w_1. So $\neg O(accept)$ and indeed, $O(\neg accept)$. For we have seen that the nearest $\neg$*accept*-world is @$_2$ and the nearest *accept*-world is w_1 and @$_2$ is better than w_1.

Finally consider, whether you ought to behave. We can see that the nearest *behave*-world is @$_2$ and the nearest $\neg$*behave*-world is w_4. But w_4 is better than @$_2$. So $\neg O(behave)$ and indeed, $O(\neg behave)$. Thus, DISJUNCTIVE DIVISION can fail in Goble's theory.

Nonetheless, we can show that $\mathbf{Bt}_{@_2}$ is interpolative in this case. Begin by noticing that Figure 2 is a representation of the following in-

formation[14]:

$$@_2 \leq_{@_2} w_4 \leq_{@_2} w_1 \leq_{@_2} w_3$$
$$w_4 \ \mathbf{Bt}_{@_2} w_3 \ \mathbf{Bt}_{@_2} @_2 \ \mathbf{Bt}_{@_2} w_1$$

Next replace each world in the ordering with the proposition that is true at exactly that world:

$$\neg accept \wedge behave \leq_{@_2} \neg accept \wedge \neg behave \leq_{@_2} accept \wedge \neg behave \leq_{@_2} accept \wedge behave$$
$$\neg accept \wedge \neg behave \ \mathbf{Bt}_{@_2} accept \wedge behave \ \mathbf{Bt}_{@_2} \neg accept \wedge behave \ \mathbf{Bt}_{@_2} accept \wedge \neg behave$$

We extract an ordering on *accept* and *behave* and their negations, by placing them in the same spot in the betterness ordering as the closest world in which they hold. We do this because Goble's semantics determines the deontic status of these claims by considering the value of the closest worlds in which they hold. So this gets us the following richer ordering:

$$\neg accept \wedge \neg behave, \neg behave$$
$$\mathbf{Bt}_{@_2}$$
$$accept \wedge behave$$
$$\mathbf{Bt}_{@_2}$$
$$\neg accept \wedge behave, \neg accept, behave$$
$$\mathbf{Bt}_{@_2}$$
$$accept \wedge \neg behave, accept$$

We similarly extract an ordering on disjunction by placing them in the same spot in the ordering as the closest world in which they hold:

$$\neg accept \wedge \neg behave, \neg behave, accept \vee \neg behave$$
$$\mathbf{Bt}_{@_2}$$
$$accept \wedge behave$$
$$\mathbf{Bt}_{@_2}$$
$$\neg accept \wedge behave, \neg accept, behave, accept \vee behave, \neg accept \vee behave, \neg accept \vee \neg behave$$
$$\mathbf{Bt}_{@_2}$$
$$accept \wedge \neg behave, accept$$

[14]Beware in translating between Goble's and Hansson's theories: Hansson's $\geq$ is most similar to Goble's $\mathbf{Bt}_w$ rather than Goble's $\leq_w$!

Notice that this ordering is in fact interpolative in Hansson's senses. Each disjunction is not strictly preferred to both of its disjuncts and is not strictly dispreferred to each of its disjuncts.

So why the different results about what we ought to do? The crucial difference is not the structure of the ordering. It is, instead, the decision rule one uses to determine what is obligatory given the ordering. On Hansson's theory, there is a fixed spot in the ordering and one checks to see if $\neg p$ is below that spot. If so, p is obligatory. But there is no fixed spot such as this in Goble's theory. Instead for each proposition p, one goes to p's spot in the ordering and checks to see if $\neg p$ is below that spot. If so, p is obligatory. This difference in the views is what explains why Goble's theory fails to validate DISJUNCTIVE DISTRIBUTION.

Since we are assuming in this example you ought to accept and behave, we know the threshold in the ordering for Hansson will have to be between the row for $accept \wedge behave$ and its negation $\neg accept \vee \neg behave$. Using this threshold, we can see one also ought to *accept* because $\neg accept$ is below the threshold so DISJUNCTIVE DIVISION holds. Interestingly, in this example, Hanssons's theory says that one ought to accept (as we just saw) and that one ought to not accept (accept is at the lowest spot in the ordering so below the threshold).

As we noted earlier, Goble's theory validates NO STRICT CONFLICTS so does not allow for cases where one ought to accept and one ought to not accept. And this is precisely because the threshold on his theory is relative to each proposition. For *accept* one compares it to $\neg accept$ and vice versa, the result is that one ought to not accept and it is not the case that one ought to accept. For $accept \wedge behave$ one compares it to its negation and similarly for *behave* and $\neg behave$. In each of these cases the threshold varies being as low as the fourth row (for *accept*) and as high as the first row (for $\neg behave$).

We will discuss issues related to distinct decision rules in greater detail below (§2.3.1). But what should the consequentialist and actualist make of these two different approaches to this variant of Zimmerman's wedding case? Certainly standard act consequentialism is more similar to Goble's theory.

Hansson's theory is however more similar to a consequentialist theories that are sometimes called absolute level satisficing theories. According to these theories, there is a some level of goodness such that if an act produces at least that much goodness, it is permissible.[15] Hansson's

[15]See [Hurka, 1990] and [Slote, 1984] for developments of satisficing approaches; see [Bradley, 2006] for criticism

theory is different from this one in two ways. First, one determines p's deontic status on Hansson's theory not by looking at p's place in the ordering but instead by looking at $\neg p$'s place in the ordering. Second, Hansson uses the ordering to determine obligations rather than permissions as the satisficing absolute level consequentialist does. Nonetheless, they are united in thinking there is some threshold. And that threshold determines the forbidden. One interesting result of our discussion is that a further data point on which consequentialist of the standard sort and absolute level satisficing consequentialists may disagree is the status of DISJUNCTIVE DIVISION.

In sum, both Goble and Hansson develop logics that can coherently model actualists and consequentialist reasoning. These logic differ about, among other things, whether they validate DISJUNCTIVE DIVISION and whether they validate NO STRICT CONFLICTS. This in turn is related to a difference in how goodness determines what is obligatory according to each theory, a difference in decision rule. So these logics are models of distinct actualist theories: Goble's logic fits best with standard consequentialism while Hanssons's perhaps fits best with absolute level satisficing consequentialist theories.

2.1.3 Semantics with Agency

As we have seen, the debate between actualists and possibilists concerns the relevance of one's future actions to what one presently ought to do. But so far, we have not considered any logics that explicitly represent the decisions and actions of agents through time. We turn to considering such theories now. We will look at the logic developed by John Horty in his [Horty, 2001] as it is perhaps the most well-known such system.

Horty's theory can be thought of as combining a certain logic of agency, so-called *stit* logic, and a certain kind of preference semantics. We begin by introducing *stit* logic.

stit logic, in turn, can be thought of as beginning with a logic of branching time and adding some way of representing agency within this framework. The logic of branching time is a kind of modal logic where we have set of points, $Tree$, and we call elements of $Tree$ *moments*. We also have an ordering on moments $<$ that is irreflexive, transitive, and *tree-like* in the sense that if $m_1 < m_3$ and $m_2 < m_3$, then $m_1 < m_2$ or $m_2 < m_1$ or $m_1 = m_2$. These properties of the ordering on moments ensure that we can represent them in a way that looks like a tree: there is a single trunk and then branches emerging from this trunk. This represents the openness of the future (and the determinateness

of the past). A collection of moments is *linerally ordered* just in case for any two of them, m_1 and m_2, $m_1 < m_2$ or $m_2 < m_1$ or $m_1 = m_2$. A collections of moments is *maximally linearly ordered* just in case it is linearly ordered and not a proper subset of any linearly ordered set. Such maximally linearly ordered collections of moments are called *histories*. Intuitively, a history represent one complete way the world could develop. Using these resources to give us our frames, one can define a model and give conditions for the truth of various claims about the past and the future. It turns out to be best in this setting to define the truth of formulas relative to a pair consisting of a moment and a history through that moment. We skip the details of the semantics for the temporal logic as it will not be of central interest in what is to come.

To this temporal logic, we add resources for representing agency. We add a set *Agent* that is thought of as populated by the individual agents that are of interest to us. We also add a function *Choice* that takes one from an agent and a moment to a partition of the histories though that moment. The cells of the partition are intuitively thought of as the actions that we can perform at that moment. Our acts allow us to select among the histories through a moment which collection of these history will occur. We also say *Choice* applied to an agent, moment, *and a history through that moment* returns whatever act that history belongs to. This represent the act that an agent does at a moment/history pair.

We now introduce in addition to the normal kinds of sentences one has, sentences concerning what an agent does or claims saying that an agent "sees to it that". There are a variety of such operators that have been proposed but we focus on a simple one often called *cstit* because it is a "sees to it that"-operator that is closely related to some ideas originally due to Brian Chellas ([Chellas, 1969]). Intuitively, the idea is that an agent sees to it that A at a moment and history through that moment, just in case A is guaranteed to be true by the act that the agent does at at that moment/history pair.

Putting all of this together, we have a *stit frame*, $\langle Tree, <, Agent, Choice\rangle$. A model M based on the frame gives the truth values of formulas at pairs, m/h, of moments and histories through those moments in the usual way for standard formulas with the following truth conditions for the special *cstit* operator:

> $[\alpha$ *cstit:* $A]$ is true relative to $M, m/h$ iff $Choice^{m,\alpha}(h) \subseteq |A|^{M,m}$

where $Choice^{m,\alpha}(h)$ represents the act α does at the moment and history through that moment pair and $|A|^{M,m}$ is the set of histories through m

such that A is true at the pair consisting of m and that history. So this just says, as intended, that α sees to it that A at a moment and history through that moment, m/h, just in case the act the agent does at m/h ensures the truth of A. Within a framework like this one, one can study the agency individuals as well as groups (a topic we return to in §4.1.2) including modeling claims about what people or groups are able to do.

Though there is much to say about this, we turn now to how one can add a deontic logic on top of this framework. The simplest way to do this would be to simply include an additional function that maps each moment into a set of histories through that moment which are ideal. While this may be a satisfactory account for what things ought to be, Horty argues at length that it is not a satisfying account of what we ought to do.[16] We do not have space here to consider these arguments in detail. Instead, I will simply state Horty's preferred approach.[17]

Horty's approach is to work with a dominance ordering on actions. He constructs this ordering by assuming there is a function $Value$ that assigns to each history a number that is understood to be a measure of how good that history is. He then lifts this ordering on histories to be an ordering on arbitrary propositions We write $P \leq Q$ to mean 'P is weakly preferred to Q'. And $P \leq Q$ iff $Value(h) \leq Value(h')$ for all $h \in P$ and $h' \in Q$. We use $P < Q$ then in the standard way to mean that $P \leq Q$ and $Q \not\leq P$.

Finally, Horty uses this preference ordering on propositions to construct a dominance ordering on act. The idea Horty's weak dominance ordering is intends to capture is the idea of one act being at least as preferred to another in every given state of the world that is independent of the act. These states of the world that are independent of the act are just understood to be given by the choice set of all the other agents (where we may include mother nature as one agent as well). So $State^{m,\alpha}$ is introduced and understood to be $Choice^{m,Agent-\{\alpha\}}$. One then defines the weak dominance ordering $\preceq$ on acts, K_1 and K_2, by saying $K_1 \preceq K_2$ iff $K_1 \cap S \preceq K_2 \cap S$ for each $S \in State^{m,\alpha}$. Strict dominance is understood in the usual way then so $K_1 \prec K_2$ iff $K_1 \preceq K_2$ and $K_2 \not\preceq K_1$.

This gives us our dominance ordering on acts that in turn allow us to define what it is that we ought to do. We do this via first identifying the optimal act for an agent at a moment. This is understood to be those acts available to the agent at that moment that are not strictly

[16] See [Horty, 2001]: ch. 3 for discussion.

[17] But see §5.1 for an overview of the issues Horty and others are responding to.

dominated:

> $Optimal^{m,\alpha} = \{K \in Choice^{m,\alpha} \mid$ there is no $K' \in Choice^{m,\alpha}$ such that $K \prec K'\}$.

Finally, then, we introduce the two place operator $\odot[\ldots cstit \ldots]$ and we form sentence $\odot[\alpha\ cstit:\ A]$ that we read as 'α ought to see to it that A' and is true just in case every optimal act ensures the truth of A:

> $\odot[\alpha\ cstit:\ A]$ is true relative to $M, m/h$ iff $K \subseteq |A|^{M,m}$ for each $K \in Optimal^{m,\alpha}$

where we assume our models are now based on frames enriched with $Value$ and the other material we have defined in terms of these items.[18][19]

So what does this theory have to say about our target cases? We now, as is shown in Figure 3, have the resources to model the case as one where the agent first faces a choice of whether to accept or not and later if one accepts faces a choice to behave or not. The idea here is at m_1,

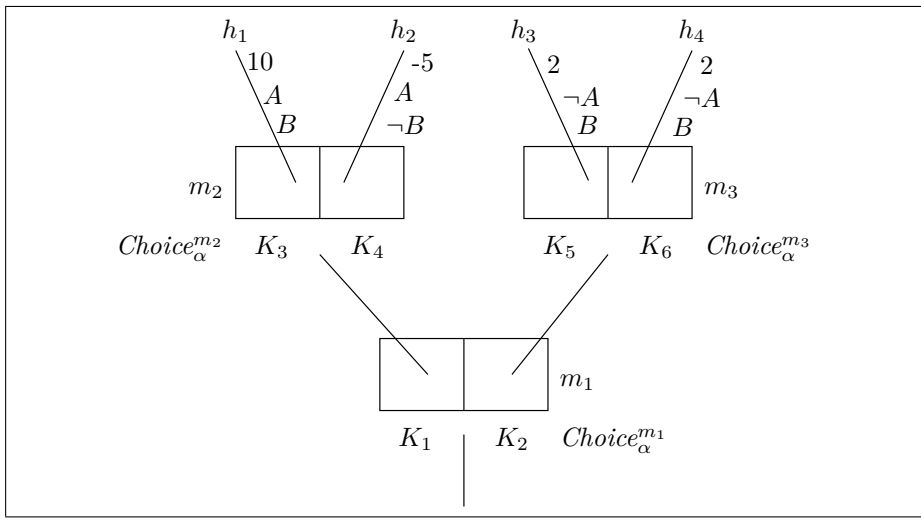


Figure 3: Wedding attendance decision tree

one chooses whether to accept the wedding invitation. K_1 represents

[18]This definition is actually not Horty's. It is however a theorem of Horty's theory that holds in finite settings. The official definition Horty gives is more complicated and allows him to handle cases that arise in certain infinite settings. Since we ignore those cases here, I opt to use as a definition of the operator what is a theorem in Horty's system.

[19]Horty also develops an obligation operator that is closer to standard act consequentialism in [Horty, 2001, §5.4]. We do not have space to explicitly present and discuss this analysis here.

the choice of accepting the invitation. Accordingly, A is the proposition that you accept and it is true in every history that you get to if you do K_1. K_2 on the other hand is not accepting so $\neg A$ is true in all of those histories. If you choose to accept the invitation, you arrive at m_2 and faces a choice of whether to behave. K_3 is the act of behaving and accordingly B which represents the claim that you behave is true in the history where you do K_3. Since this is a very good outcome, we have given the history the value of 10. On the other hand, one could choose K_4 and not behave. Accordingly, $\neg B$ is true here and it is assigned a low value of -5. On the other hand, at m_1 you could choose to not accept. This would the result in you arriving at m_3 where you can choose between K_5 and K_6 which are perhaps simply different ways of spending sometime like reading a book or watching tv. Regardless of what you do, you will have behave. Accordingly, B is true in these histories and each history is of some small middling value 2.

What does Horty's theory claim that you ought to do in this case? More precisely, at m_1 what should one do. To determine, this we start with the optimal acts. They will come from the set $Choice^{m_1,\alpha}$ so we need to check to see how K_1 and K_2 rank relative to one another. In situations where, the agent is the only one choosing such as this one the set $State^{m_1,\alpha}$ will only contain one element which is the set of all histories through m_1, $\{h_1, h_2, h_3, h_4\}$. Accordingly, the dominance ordering and the preference ordering will be the same. Since we know $Value(h_1) > Value(h_3)$ and $Value(h_2) < Value(h_3)$, we know neither K_1 nor K_2 are weakly preferred to one another so there is no dominance. Thus, both K_1 and K_2 are non-dominated so both are optimal.

In this setting, $\odot[\alpha$ *cstit:* $A \wedge B]$, $\odot[\alpha$ *cstit:* $A]$, and $\odot[\alpha$ *cstit:* $\neg A]$ are all false. $A \wedge B$ is not guaranteed to be true by K_1 and it is guaranteed to be false by K_2. A is not guaranteed to be true by K_2. $\neg A$ is not guaranteed to be true by K_1. Thus, there is nothing that one ought to do in this case. Instead, one is merely permitted to do any of these things.

This is interesting as it corresponds to neither the actualist nor the possibilist view. It is however like the actualist view in that it does not treat your future decisions differently than simply states of the world that you don't control. To see this, we can just notice that the very same results follow if we simply have no choices at m_2 and m_3. It however differs from standard consequentialist actualist views in that it does not assume a single fully specified possibility results from one's act. We discuss this abstract difference between standard consequentialism, Horty's theory, and the other theories that we have described in §2.3.1.

But one strange result of Horty's theory is that even the claim $\odot[\alpha$ *cstit:* $\neg(A \wedge \neg B)]$ is false. While K_2 guarantees $\neg(A \wedge \neg B)$, K_1 does not guarantee the truth of this as it is false at h_2. Yet it seems hard to deny that one ought to not both accept and misbehave. So it appears that this is not a sensible treatment of our case.

Perhaps unsurprisingly then, Horty does not himself endorse this way of using his theory to analyze this example.[20] Instead, Horty expands his theory to deal with cases in which one faces a sequence of choices. The expansion involves introducing the notion of a strategy. It turns out to be a somewhat delicate matter how a strategy must be defined and how a "strategic"-ought-to-do operator, $\odot[\ldots \mathit{scstit} \ldots]$ is to be defined. But we will pass over these complications and work with a simplified intuitive picture as this will be enough for our example.

In our example, the available strategies are these:

$$s_1 = \{\langle m_1, K_1\rangle, \langle m_2, K_3\rangle\},\ s_2 = \{\langle m_1, K_1\rangle, \langle m_2, K_4\rangle\},$$
$$s_3 = \{\langle m_1, K_2\rangle, \langle m_3, K_5\rangle\},\ s_4 = \{\langle m_1, K_2\rangle, \langle m_3, K_5\rangle\}$$

Each strategy, then, is a coherent sequence of choices one could make at a moment. Roughly, one considers which strategy is optimal by (in a setting where the state is just all of the histories) checking whether every history in one is at least as good as every history in the other. It is easy to see this reduces to comparing the value of h_1, h_2, h_3, and h_4 and obviously h_1 is higher than all of them so s_1 will strictly dominate all of them and so be the only optimal strategy. From here one evaluates $\odot[\alpha$ *scstit:* $A \wedge B]$, $\odot[\alpha$ *scstit:* $A]$, $\odot[\alpha$ *scstit:* $\neg A]$, etc. by considering whether s_1 settles the embedded sentence. And it is easy to see s_1 ensures the truth of $A \wedge B$ and A and ensures the falsity of $\neg A$. Thus, this version of Horty's view agrees with the possibilist

So on Horty's picture there is a way of accommodating possibilist reasoning about these cases. But, as Horty himself notes, there is no obvious way of accommodating actualist reasoning in his theory.[21]

In sum then, Horty's picture gives us two different views about what we ought to do in this case depending on whether focus on the $\odot[\ldots \mathit{cstit} \ldots]$ or the $\odot[\ldots \mathit{scstit} \ldots]$ operator. And the view we get about the $\odot[\ldots \mathit{cstit} \ldots]$ operator corresponds to no major view in ethical theory while the $\odot[\ldots \mathit{scstit} \ldots]$ has similarities to possibilist views.

[20] Horty's discussion of the actualism/possibilism debate is found in [Horty, 2001, §7.4.3].

[21] Horty does suggest that the actualist claim may correspond to the claim that one ought to not accept conditional on not behaving. But this claim is uncontroversial and not something the possibilist would disagree with this.

We face, then, a question of what role these different operators play and whether one operator is to be privileged above another operator. We also face a question of whether any plausible view in ethics corresponds to the claims made about our example by the $\odot[\ldots cstit \ldots]$ operator.

Thus, while Horty's view allows us to more explicitly spell out the role of diachronic agency in determining what we are obligated to do and allows to provide a model for coherent possibilist reasoning, a number of questions remain about how to best understand this view. In the next subsection, I present a leading idea in ethics about the actualism and possibilism debate and suggest that it helps to shed new light on some of the issues for Horty's theory as well as the issues for Goble and Hansson's theory that I mentioned earlier.

2.2 An Insight From Ethics

We turn now to a prominent line of criticism of the simple form of actualism that I have presented and line of defending something akin to DEONTIC INHERITANCE from ethical theory.[22] These ideas originate in the work of Holly Smith and have since been developed further by Douglas Portmore and Jacob Ross.[23] In this article, we will only present a simplified version of some of the basics of this approach.

The line of thought involves a focus on the details of the connections between agency and what we ought to do. To get a feel for the view, it helps to begin with a case everyone agrees on. Suppose the best thing for me to do would be to push this button before me. It would, suppose, result in everyone being slightly happier. Suppose further that I easily can push the button: it is right in front of me, I see it, my arms are in good working order. But suppose that I will not push the button because I do not intend to do so currently. Perhaps, this is because I am a cruel person and so have decided to not push the button. In this setting, the fact that I will not push the button is simply not relevant to whether I ought to push it.

Why is that? The idea, I take it, is straightforward enough: since it is up to me whether I push the button the fact that I will not push it is

[22]Strictly speaking, these approaches defend a restricted version of inheritance. Very roughly, inheritance is restricted to applying to cases where doing one act involves doing another act. The details of how to spell this require a better understanding of the notion of "what is up to me" that is invoked below. This is an area where some formal work could also benefit those in ethics as the notion of an act "involving" another act could do with some formal clarification.

[23]See [Goldman, 1978; Portmore, 2011; Portmore, 2013; Ross, 2012]. These ideas are also related to so-called "maximalism", see [Brown, 2018] and [Portmore, 2019].

no excuse. Now there is a question of in what way is it "up to me" and different theories can make this notion more precise in different ways. But a simple idea is that it is up to me in that is under the control of my present intentions where we say:

> S's doing x is under S's present intentional control at t iff if S were to intend at t to do x, S would do x and if S were to intend at t to do $\neg x$, S would do $\neg x$

The idea is that facts that are under our present intentional control can't get us off the hook for doing things that would bring about what is best.

This idea can be applied to the wedding case. There are, we can suppose, two different interpretations of this case. On one interpretation, if I were to intend now to accept the invitation and behave, I would in fact end up accepting the invitation and behaving at the wedding. Now of course, I do not in fact have this intention even though I could have it. Instead, I have no intention at all with regard to whether I will behave. For this reason, if I were to accept (and indeed even if I were to intend to accept and in fact accepted), I would not behave myself. On this interpretation, the idea is that I ought to accept the invitation even though accepting it would lead to a worse outcome. This is because, the view says, we should not hold fixed the fact that I will misbehave in evaluating whether to accept. We should not hold it fixed because it is under my present intentional control whether I misbehave.

On another interpretation, even if I were to intend now to accept the invitation and behave, I would not end up behaving. In this setting, whether I behave is not under my present intentional control (though, on the natural way of understanding the case, it is at the time of the wedding under my intentional control whether I behave). And the idea here is that I ought not to accept the invitation. Nonetheless, this verdict, plausibly, is compatible with DEONTIC INHERITANCE. This is because it is plausible that if it is not up to me at the present moment whether I accept and behave, then I cannot accept and behave in the sense of 'can' that is featured in the 'ought' implies 'can' principle. For this reason, it is plausible that it is not the case that I ought to accept and behave and so DEONTIC INHERITANCE is not violated.

Now there are many important questions and worries about this position. And there are a number of variants of this position that deal with the interesting issues about how to make sense of the notion of control or the notion of what is up to me precise and plausible.[24] But

[24]See especially [Portmore, 2019].

even without diving into the details of this rich topic, there are a number of lessons that we can learn from this insight from moral philosophy.

First, we noted earlier (§2.1.2) that views like Lou Goble's consequentialist inspired theory make what is obligatory dependent on what one will actually do in a somewhat curious way. This idea in moral philosophy identifies more precisely what the problematic dependency might be and provides a compelling criticism of it: while the relevant states of the world for determining what we are obligated to do may be partially determined by facts about what will (or would) happen, they do not depend on those facts that are under one's intentional control.

Second, this insight from moral philosophy puts at center stage connection between features of our agency and our obligations. This suggest that a theory like Horty's theory that involves representations of features of agency that are relevant to determining what agents ought to do may be closely related to this idea in moral philosophy.

But in fact, at least on first inspection, it does not seem as those Horty's theory makes use of the notion of intentional control or can help us to get a better sense of it. Horty's theory, instead, is built with a much sparser and better understood (albeit perhaps more idealized) set of resources than the ideas that I have been rehearsing here. That said, I believe we can use these ideas in moral philosophy to provide a certain helpful interpretation of some features of Horty's theory.

To see this, recall that we left our discussion of Horty's theory with two questions. First, what should we make of the fact that the $\odot[\ldots cstit \ldots]$ approach to the case did not correspond any sensible interpretation of it for it led to results such as $\neg \odot [\alpha \ cstit{:}\ \neg(A \wedge \neg B)]$. Second, while the $\odot[\ldots scstit \ldots]$ account was a natural development of the possibilist line of thought, it is not clear when this operator is the appropriate one to guide our action as opposed to the simpler $\odot[\ldots cstit \ldots]$ operator. In other words, it is not clear what different theoretical and practical roles these operators are to play and how they are related.

Now Horty does in fact quite precisely describe a certain kind of relationship between the operators. In the set up to discussing the strategic-cstit operator, Horty points out that often we do not want to consider every single possible future decisions but instead only decisions up to a time. He implements this in his models by adding a parameter for what he calls a "field of concern". Generally a field of concern is some subset of the total histories and strategic cstit is defined only over strategies constructible in the field of concern. Horty shows that if the field of concern is only this very moment $\odot[\ldots cstit \ldots]$ and $\odot[\ldots scstit \ldots]$ are

equivalent.

But what field of concern is appropriate for a given problem? This is a question Horty does not answer. Perhaps, this is because there is no fixed answer or perhaps because it is a context-sensitive issue. But one suggestion that I would make is that the view from moral philosophy that I have just described gives us grounds for thinking that there is a privileged field of concern for each question about what one ought to do at a given moment.[25] It is that space of choices that is currently under your intentional control. If this is right, it suggests that the only relevant operator for modelling the core moral claims that we would like to make is the $\odot[\ldots \mathit{scstit} \ldots]$ operator evaluated relative to this privileged field of concern. This eliminates the question of which operator to give priority.

Following this line of thought, we can consider each interpretation of the case. The first interpretation recall says that one would accept the invitation and behave if one (presently) intended to accept the invitation and behave. So the field of concern is not just the present moment but includes moments after this. This explains why $\odot[\ldots \mathit{cstit} \ldots]$ doesn't correspond to any sensible interpretation of the case: it is inappropriate to have only this moment is your field of concern when your intentional control extends beyond this moment. And we, as we saw earlier, have the possibilist result that you ought to accept and behave and that you ought to accept.

If, on the other hand, we interpret the case so that one does not have present intentional control over whether one will behaves, the field of concern diminishes to the present moment. In this context, we saw that it is not the case that one ought to accept and behave because one cannot do this. Indeed, this follows from theorem of Horty's view that says that $\neg\Diamond[\alpha \textit{ scstit: } A] \rightarrow \neg \odot [\alpha \textit{ scstit: } A]$. What's more on this way of thinking, it make sense why: since one is unable to accept and behave and one is unable to accept and not behave, this cannot be something one is obligated to do.

However, one discrepancy remains between the idea that I have been discussing from ethics and this interpretation of Horty's view. According to the idea in ethics, one ought to reject in the situation where one does

[25]One thing to notice is that $\mathit{Choice}^{m,\alpha}$ is a partition of the histories through that moment where each "act", K, is cell of the partition. This is a kind of maximalism (of the sort discussed in n. 23) limited acts at a moment. Once we expand to discuss strategies, the resulting choice set also forms a partition that is typically more fine-grained and in this way represents a maximalism that is consonant with the fact that the maximal acts that we perform often extend over time.

not have present intentional control over whether one will behave. This is not so on Horty's view implemented in the way that I have described. According to this view, one is permitted but not required to reject.[26]

What account for this difference is that the idea from ethics assumes that a single possibilities results from what one does. And in cases where this possibility is not determined by facts that are under one's present intentional control, this possibility is the only relevant outcome of the act. This is not so on Horty's theory. On Horty's theory, any future action available to the agent (even if it is outside the field of concern) makes it so there are multiple potential outcomes of the act. As we will see, this abstract difference between the the theories has a number of broader consequences and it is also a difference between Horty's theory and the theories that we have discussed before this. We will explore these more abstract differences as well as other issues below (§2.3.1).

For now though, let us take stock. The theories that we have discussed in this previous subsection (§2.1) allowed for both actualist, possibilist, and still other verdicts that are not easy to categorize. While some of the theories, such as SDL, offered us little insight as to why we should expect one set of verdicts or one form of reasoning to be correct, the theories of Goble, Hansson, and Horty offered more insight. This shows some of the fruits of formal theorizing for issues in ethics: it allows us to more fully and systematically explore the range of possible views and it allows to trace the differences between certain verdicts to different views about how rankings on outcomes are related to what we are obligated to do.

We have also seen in this subsection that ideas from ethics can help us to understand certain features of the theories from logic that are harder to grasp in the abstract. In particular, we noticed that we can get a better grip about what is strange about certain dependence of the obligatory on what will be done and we can get a better understanding of which future potential choice opportunities are relevant to our present obligation. And there is every reason to be optimistic that more focused attention than can be provided in a handbook article will yield still more results.

2.3 Further Issues

We close our discussion in this section by considering some further features of modal deontic logic that are relevant to ethical theory.

[26] We also continue to have the seemingly strange result that $\neg \odot [\alpha \textit{ scstit: } \neg(A \wedge \neg B)]$ even though $\Diamond[\alpha \textit{ scstit: } \neg(A \wedge \neg B)]$.

2.3.1 Decision Rules in Modal Semantics

We start with a certain high level difference between these theories. This difference is relevant to the treatment of actualism/possibiiism debate. For example, we touched on it in comparing the theories of Goble and Hansson (§2.1.2). But the difference is so general that it has effects that percolate down to the verdicts that the theories give about many kinds of cases. The heart of the issue is a certain push and pull between two structural features of these theories. The first features concerns which decision rule to adopt for selecting among acts, outcomes, etc. and how fine-grained that decision rule should be. The second structural feature concerns which possibilites are to count as the "relevant outcomes" for determining what is obligatory.

We begin our discussion of this by returning to Hansson's theory. Above, we only discussed how Hansson's theory ordered propositions. But Hansson also mentions how relations among proposition may be related to relations among world. These ideas are summarized in Table 1 (from [Hansson, 2013, 492]) where max is a function that takes us from a proposition to the best world in which that proposition is true and min is a function that takes us from a proposition to the worst world in which it is true.

Maximin preferences:	
$p \geq_i q$ iff	$min(p) \geq min(q)$
Maximax preferences:	
$p \geq_x q$ iff	$max(p) \geq max(q)$
Interval maximin preferences:	
$p \geq_{ix} q$ iff	either $min(p) > min(q)$
	or both $min(p) \simeq min(q)$ and $max(p) \geq max(q)$
Interval maximin preferences:	
$p \geq_{xi} q$ iff	either $max(p) > max(q)$
	or both $max(p) \simeq max(q)$ and $min(p) \geq min(q)$
Doubly maximizing preferences:	
$p \geq_{\ddagger} q$ iff	$max(p) \geq max(q)$ and $min(p) \geq min(q)$

Table 1: Possible relations between orderings on propositions and orderings on worlds in Hansson's theory

One way to see what is interesting about these connections is that they enforce a certain kind of decision rule or, more accurately, an elimination or negation selection rule. To see this, recall that $O(p)$ just in

case $f \geq \neg p$. So whether $\neg p$ is eliminated and p is adopted is related to the underlying ordering of worlds. To give one example, consider *Maximin preferences*. If the worst $\neg p$ worlds are worse than or equal to the worst f worlds, then $f \geq \neg p$ and thus $F(\neg p)$ and $O(p)$. So $\neg p$ can be eliminated and p can be adopted. So the maximin criteria gives us an elimination/negation selection rule.

Hansson, for his part, prefers *Doubly maximizing preferences* because this insures that the relation among propositions is interpolative. However formally appealing this might be, in many settings the resulting ordering on propositions will be incomplete or tied as the decision rules need not strictly order all incompatible propositions. In cases like this (e.g., any pair of gambles where one has a greater net pay out if you win and also greater net loss if you lose), this will result in both options being permissible even though intuitively they may not both be permissible (e.g., when one option has greater expected value).

Similar points can be made about the other options in Table 1. And indeed, they extend even to the other theories that we have considered. Let's take a look at this.

To simplify matters, let us consider how these different theories determine whether it is obligatory that p or obligatory that $\neg p$ or neither. SDL does this by simply checking a set of worlds to sees if they unanimously say p, or unanimously say $\neg p$, or give no unanimous answer. As a decision rule then, SDL implicitly suggest that if p is obligatory, then it passes a maxi-max test (the best p-worlds are better than the best $\neg p$-world). The other theories involve more explicit comparisons. Goble adopts a Pareto-like rule so that p is ranked ahead of $\neg p$ just in case every p-world is as good as every $\neg p$-world and at least one is better. Horty's theory is harder to succinctly describe as it involves ranking on specific histories, rankings on propositions (sets of histories), and rankings on acts. But Horty also adopt certain qualitative rules. He adopts a Pareto-like rule for determining the ranking on propositions and then a state-wise dominance rule to determine rankings on acts.

What all of these theories have in common is that they adopt qualitative decision rules that do not appeal to any ideas about the likelihood of outcomes. What's more, many of them also do not appeal to the idea that the goodness of possibilities is numerically measurable. This makes it difficult for these theories to get the result that something is obligatory in any case in which the spread of relevant outcomes for an act is sufficiently wide and diverse in terms of goodness as compared to their competing outcomes.

Strikingly, none of these proposals adopt rules of the sort familiar

from expected value decision theory, rules that make use of a weighted average determined by the numerically measurable values weighted by probabilities. Of course, in order for expected value approach to work, we need some sensible way to numerically measure the values and some way to interpret what a probability is. And whatever procedure we adopt for this we need to make sense of how to assign these quantities to worlds, propositions, and acts, and why these quantities can be sensibly multiplied and added. There is a rich philosophical literature on the "additivity" of value and a rich formal literature on the measurability of quantities that can contribute to this.[27] In certain cases, the issue is well-understood. For example, if we interpret our ordering as telling us an agent's subjective preferences, decision theorists have representation theorems that tell us what conditions that ordering must satisfy in order for us to make sense of taking a weighted average.[28] Other cases are less well-understood, however.

For reasons like this and others, the theories that we have discussed have avoid using these kinds of expected value decisions rules. Here is what Horty and Hanssons have to say:

> The particular ordering that results from comparison of expected value relies, however on a kind of probabilistic information concerning outcomes of actions that is often either unavailable or meaningless; and this is true especially in situations in which the outcome resulting from an agent's action may depend, not simply on a roll of the dice, but on the independent choice of another free agent.[Horty, 2001, 59–60]

> However, this is not suitable explication of preferences for the purpose of deontic logic. Suppose that you are deliberating on whether to keep an extra income for yourself (s) or donate it to a charity (c). The probability that you will do one or the other [...] should not influence your choice since if it did, then you would not really treat both alternatives as fully open. [...] In addition, there is a counter-argument of a more formal nature: Weighted-average preferences are not in general interpolative and they do not even satisfy the highly plausible property that if $p \simeq q$, then $p \simeq p \vee q$, where $\simeq$ denotes indifference.[Hansson, 2013, 491]

[27] See [Krantz *et al.*, 2007] for a classic introduction to the topic of measurement

[28] See [Ramsey, 1931 1923] and [Jeffrey, 1990] for classic discussions; see [Meachem and Weisberg, 2011] for recent philosophical reflection on how these results are to be interpreted.

Here is not the place to assess this dispute between expected value decision theories and the theories discussed in this section. But this topic is worth further investigation by moral philosophers and deontic logicians.[29]

That said, there may in fact be a way of avoiding these difficulties. The difficulties arise because we are connecting a proposition's or act's place in the ranking with the places of various worlds where the proposition is true or the act is performed. The decision rules that we have looked at are either not strong enough to give us rankings that are as discriminating as we'd like them to be or are like the weighted average rule which requires stronger commitments than we might like. The trouble, at root, stems from the fact that there are typically many worlds in which the proposition is true or the act is performed and then having to use this multitude of worlds to determine the place of the proposition or act in the ranking.

One attractive feature of views that appeal to certain counterfactuals to determine the outcomes that are relevant to whether an act is obligatory is that they only require us to consider a (typically proper) subset of the worlds where the proposition is true or the act is performed. According to these views, the world or worlds that are relevant are the ones that would result if you did the act. According to certain standard analyses of counterfactuals, this often means there is a single world that we have to look at. Or, in any case, a small family of worlds that are as close as possible to the actual world. In this setting, it may be that any of the above decision rules will give us orderings that are sufficiently discriminating. Let us compare then how some of our theories determine what the relevant outcomes are.

We already saw Hansson's theory considers the total space of possibilities and this, of course, will also be a feature of SDL. Goble's theory, Horty's theory, and traditional forms of consequentialism, on the other hand, require modal space to have a rich enough structure to accommodate counterfactuals.

We can investigate this by looking at some properties of the closeness-orderings used in the standard semantics for counterfactuals.[30] Recall

[29] I do not mean to suggest that these issues have never been considered. Indeed, Goble in [Goble, 1996] develops an expected value consequentialist approach. I do not discuss the details of this paper partially for reasons of space. But more importantly, as I see it, the question of how to integrate decision-theoretic ideas related to values and probabilities with deontic logic is wide-open even though some initial forays have been made.

[30] See [Lewis, 2001 1973] for the canonical treatment.

that a standard gloss on how one evaluates the truth of the counterfactual $p \boxrightarrow q$ at a world w is that one finds the closest p-worlds to w and checks whether q is true at throughout those worlds.[31] But there are three related proposals presented in Table 2 about the structure of this similarity ordering that influence what kinds of theorems hold in semantics of this sort.

Properties of the closeness-ordering	Theorems
WEAK CENTERING if p is true at w, then w is one of the closest p-worlds to w	COUNTERFACTUAL MODUS PONENS $(p \boxrightarrow q) \rightarrow (p \rightarrow q)$
STRONG CENTERING if p is true at w, then w is the unique closest p-world to w	CONJUNCTION CONDITIONALIZATION $(p \wedge q) \rightarrow (p \boxrightarrow q)$
UNIQUENESS there is a w' that is the unique closest p-world to w	CONDITIONAL EXCLUDED MIDDLE $(p \boxrightarrow q) \vee (p \boxrightarrow \neg q)$

Table 2: Correspondence between properties of the closeness-ordering and theorems about counterfactuals

As the interested reader can check for themselves if they work through the details, Horty's theory accepts only WEAK CENTERING, Goble's theory in addition accepts STRONG CENTERING, and standard forms of consequentialism all three claims. How is this relevant to what we obligated to do? According to these theories, we determine whether we are obligated to do something in part by the value of various worlds. But which worlds we look at is determined by the similarity ordering. In all of these theories, in cases where I end up performing some act, the value of the actual world is relevant to the acts status. In Horty's theory, the actual world is one of perhaps several worlds values who is relevant. In Goble's theory and traditional consequentialism, it is the only world whose value is relevant In Horty and Goble's theory, we compare the value an act that you will perform in this world with the value of an incompatible acts, there are a number of worlds whose value is relevant. But in traditional consequentialism, we compare another single world.

As we know, SDL and Hansson's theory differ from this. But Hansson's theory "looks" similar to how Horty's theory works in that a non-

[31] Or at least this is how we interpret the counterfactual in settings where we assume the limit assumption is satisfied.

singleton set of worlds determine the value of p and it is compared with the value of other acts that are determined by another non-singleton set of worlds. But, in Horty's theory these sets are typically proper subsets of the set of worlds where p is true while in Hansson's theory the set of worlds is all the worlds in which p is true.

The upshot of of all of this is that our theories give us different "thinning outs" of the space of relevant outcomes. Standard consequentialism is the most extreme. On virtue of this is that there is no longer a wide and diverse set of outcomes that will lead to incompleteness or ties in the rankings of acts. Indeed, given that we are comparing two worlds as the consequentialist does, all of the decision rules in Table 1 give the same verdict about how the worlds rank. As such, consequentialism will be decided about what is obligatory in any case in which the world's are not exactly as valuable as one another. This makes it so even qualitative decision rules give a rich set of obligations.

It is harder to say in any systematic way what the result of Horty's and Goble's theory are, but we know they represent a thinning out of what the relevant outcomes are and, as such, may and likely will result in more verdicts about what is obligatory than other theories.

It is not my aim to argue that these approaches that use counterfactuals are correct and other approaches are not. All I wish to do is to bring into focus the underlying push and pull between the two structural assumptions that I identified. A theory's decision rule together the modal structure it posits determines how discriminating the ordering on acts is. On one extreme, we have extremely strict uniqueness validating modal structure posited by consequentialism which gives us a very discriminating ordering on acts according to to any decision rule. On the other extreme, we have the unrestricted modal structure together with expected value decision rule which also gives us a very discriminating ordering on acts. In between we have decision rules like maximin, Pareto, etc. and modal structures like weak centering, strong centering, etc. Moral philosophers and logicians alike can fruitully interact by considering which of these is appropriate for which applications.

In moral philosophy some of these issues have already been explored to some extent. For example, many philosophers believe there are so called "objective" obligations which do not depend on your information but just the facts and so-called "subjective" obligations that do depend on your information.[32] For some of them the objective obligations are

[32]See [Schroeder, 2018, §IIB] for a recent discussion of the theoretical role of this distinction. But the distinction itself has played a role in moral philosophy for decades (though sometimes it has gone under other names).

modelled correctly by any decision rule and uniqueness (traditional consequentialist). But subjective obligation is best modelled by the expected value rule and no modal structure. These different obligations are supposed to correspond to different theoretical roles in our practice. The first is involved in standards of correctness and giving of advice. The second is involved in assessments of rationality and portioning blame. There is also an open question of which of these ought to be taken to be of central interest.[33]

There has been considerably less work done about the intermediate cases and what there role might be. This issue about the interaction between these two structural features would benefit from deeper study than I have presented here. There are important formal issues about how to make precise (and whether it can be made precise) some of the comparisons between theories that I have made here. And there are important conceptual issues about how to understand the role of different kinds of structure and whether any of them deserves to be privileged.[34]

2.3.2 Other Interpretations

The simplest interpretation of the deontic logics that we have looked at are value-based interpretations. We can understand the set of worlds that witness the truth of claims about obligation in SDL as the worlds that are best; we can understand the underlying orderings involved in Goble, Hansson, and Horty's theory as goodness or value-based orderings. While natural, these interpretations are not inevitable as we have seen. And this is good news because in moral philosophy it is, of course, controversial whether anything like a value-based framework is correct. Let us take a look at what some other possible interpretations might be.

Perhaps, the most straight forward alternative interpretation are ones that do not stray far from the value-based approach. So for example, views which take preferences or idealized preferences to determine what is right and wrong could easily provide alternative interpretations of the orderings that our theories make use of. Of course, as with the interpretation in terms of values, one must check that the logical properties of the ordering are ones that are compatible with the interpretation.

[33]See [Lord, 2017; Lord, 2018b] for a helpful presentation of the state of play as well as a defense of a distinctive view about what is of central interest.

[34]There are still further questions about which comparison class is relevant for determining obligations, the relationship between the values of acts and the values of outcomes, and which outcome in which an act occurs is relevant for determine its deontic status. These questions are preliminary explored in [Nair, 2020a].

This is not trivial. For example, all of the theories that we have looked at assume that the ordering among worlds is connected in the sense that any two worlds or propositions are ranked with respect to one another by the ordering (where we allow this may means some worlds are tied). This amounts to the claim that value comparison can always be made or comparison according to preference can always be made. While plausible, this controversial as some believe that there are incommensurable goods.[35] Whatever interpretation we choose, we must take a stand on this issue or provide a way of relaxing the framework to allow for unconnected orderings. Similar issues arise concerning other properties of the ordering such as transitivity.[36]

But putting these very general difficulties aside, there seems to be no particular reason to be suspicious of interpretations of these logics in terms of preferences. A more interesting question however is whether moral theories that are very different are compatible with the theories that we have discussed here. For example, how can leading deontological theories like Kantianism, contractualism, or Rossian pluralism interpret the logics presented here? And how can theories in the virtue ethics tradition interpret the logics presented here?

The best supported but disappointing answer to this question is that they may be compatible with these theories. For each theory, there is a somewhat trivial way of trying to show that there is license for optimism that it is compatible with the frameworks that we have been discussing. One simply takes what is permitted according to each theory and ranks it higher or ranks the worlds in which it occurs higher than that which is not permitted. If this procedure works (it would requires some care to show it works and we already saw in §2.1.1 some grounds to be cautious about this), then we can rest assured that these moral theories are compatible with the deontic logics that we have been discussing.

But there would be very little interesting about this result. The formal theories would not give us much insight into the underlying interest or structure of the moral theory. And the moral theory would not give us much of an explanation for why the particular formal structures that we are using are sensible. For example, what is interesting about contractualism of the sort Thomas Scanlon has developed is that there is a test for reasonable rejection for rules and the test does not allow for interpersonal agregation ([Scanlon, 1998]). What a good formal theory

[35] See [Chang, 1997]: Introduction (and the other essays in the volume) for a useful survey of the issues raised by incommensurability.

[36] See [Tempkin, 2015] for a book length discussion of problems for the transitivity of goodness.

would do is give a way of representing rules, representing perspective of those who may have reasons to reject or accept such rules, and then define from these a logic that allows us to study what is obligatory according to this theory. Formally, this would be an interesting task and conceptually it would offer a more precise model of contractualist reasoning that would be helpful in moral philosophy. But none of this is provided by the theories that result from the easy procedure that I described.

I believe this lesson generalizes. While non-value- or preference-based theories may be compatible with the modal theories of the sort that we have been discussing, there is very little of interest in this fact. The distinctive aspect of these moral theories are not represented in any especially illuminating why by the formal structures and the moral theory does little by way of explaining why the formal structure is a sensible one for modeling obligation.

So one important area of research is to develop formal models that are better fit for these deontological theories and theories from virtue ethics. For most of these moral theories, very little work has been done.[37] I suspect this is partially because the moral philosophers who have worked on these theories have been less interest in formally developing their theories and more interested in explaining why they are attractive alternatives to consequentialist theories. That said, the theories are discussed in informally precise prose by many leading moral philosophers and are ripe for exploration by the more formally inclined.

That said, a notable outlier are Rossian pluralist theories and the particularistic theories that arose from Ross's insights. The theories have received considerably more attention in deontic logic. The next section introduces these deontic logics.

3 Particularism

Perhaps one of the most fruitful interactions between deontic logicians and moral philosophers concerns particularism. Particularism is, very roughly, the view that there is no codifiable set of principles or norms governing the moral.[38] Morality is, as it is sometimes said, "shapeless". We begin by looking at the motivations for this view (§3.1). We then

[37] See however [Rechenauer and Roy, 2014] for a discussion of contractualist approaces and [Braham and van Hees, 2015] and [Lindner and Bentzen, 2018] for discussion of Kantian approaches.

[38] Leading discussions of particularism include [Dancy, 2004; Hooker and Little, 2001; McKeever and Ridge, 2006].

present a formal system that allows us to model many of the phenomena that motivate this view (§3.2). We close by assessing how satisfied we should be with this model (§3.3, §3.4).

3.1 Motivations

For almost any putative principle one can come up with cases where it seems not to be in force. This was famously noted by Henry Sidgwick in his critique of the intuitional method in *The Methods of Ethics* ([Sidgwick, 1981 1907]: Book 3). Sidgwick painstakingly considers various principles that, for example, forbid lying or enjoin you to keep your promises and argues that they do not apply correctly in certain cases. And Sidgwick argues that various modifications of these principles are subject to problems of their own as well.

Sidgwick concluded from this that the utilitarian principle is the best explanation of why these principles give us the correct results in some cases and fail in other case.[39] But philosophers since Sidgwick have seen utilitarianism as subject to similar problems. G.E. Moore thought the culprit was the simplistic hedonist axiology implicit in Sidgwick's utilitarianism and opted for a perfectionist form of consequentialism instead [Moore, 1962 1903, especially ch. 3 and 6]. But W. D. Ross suggested that even this was not enough. Consequentialism as such cannot accommodate all the cases. Here is one of his examples:

> If I have promised to meet a friend at a particular time for some trivial purpose, I should certainly think myself justified in breaking my engagement if by doing so I could prevent a serious accident or bring relief to the victims of one. And the supporters of the view we are examining hold that my thinking so is due to my thinking that I shall bring more good into existence by the one action than by the other. A different account may, however, be given of the matter, an account which will, I believe, show itself to be the true one. It may be said that besides the duty of fulfilling promises I have and recognize a duty of relieving distress, and that when I think it right to do the latter at the cost of not doing the former, it is not because I think I shall produce more good thereby but because I think it the duty which is in the circumstances more of a duty. This account surely

[39] Sidgwick also argued that the few intuitional principles that withstand scrutiny, in fact, entail utilitarianism.

> corresponds much more closely with what we really think in such a situation. If, so far as I can see, I could bring equal amounts of good into being by fulfilling my promise and by helping someone to whom I had made no promise, I should not hesitate to regard the former as my duty. [Ross, 1930, 18]

As the example shows, Ross's preferred view is that there are a number of prima facie duties. In a given case, one or several of them may apply. What we ought to do is determined by which of the duties is stronger in this context.

But Ross did not think that we could provide any once-and-for-all list that ranks the duties according to strength such that one ought to do what is suggested by one's strongest applicable duty. This is because Ross believed that the relative strength of duties can vary from case-to-case. He writes:

> "But no act is ever, in virtue of falling under some general description, necessarily actually right; its rightness depends on its whole nature and not on any element in it." (ibid.: 33)

> Every act therefore, viewed in some aspects, will be prima facie right, and viewed in others, prima facie wrong, and right acts can be distinguished from wrong acts only as being those which, of all those possible for the agent in the circumstances, have the greatest balance of prima facie rightness, in those respects in which they are prima facie right, over their prima facie wrongness, in those respects in which they are prima facie wrong — prima facie rightness and wrongness being understood in the sense previously explained. For the estimation of the comparative stringency of these prima facie obligations no general rules can, so far as I can see, be laid down. [Ross, 1930, 41]

So for Ross, we can state some moral principles that tell us certain considerations (e.g., promise keeping) have force in any case where they apply (e.g., any case in which an agent has made a promise, keeping the promise is prima facie right). But we cannot give a procedure or rule for going from this list of contributing factor to what we ought to do overall: while it is possible to give a principled account of *what* factors contribute to rightness, it is not possible to give a principled account of *how* they contribute. This then is one grounds for the particularist thought.

But there is an even deeper sense in which a particularist believe the morality cannot be understood in terms of principles. They reject even Ross's thought that we can give a principled account of what factors contribute to rightness. They argue for this by showing that many putative factors that contribute to rightness in one context are *undercut* in other contexts so that they have no force at all or are *intensified* so that they may more strongly contribute in some contexts or are *attenuated* so that they more weakly contribute in other contexts. Here are some examples that have been offered:

> we might point out that in some contexts the fact that something is against the law is a reason not to do it, but in others it is a reason to do it (so as to protest, let us say, against the existence of a law governing an aspect of private life with which the law should not interfere). [Dancy, 2017]

> Not only is it possible to think of cases in which it is false that one ought not to lie, it is also possible to think of cases in which it is false that the fact that some action would involve lying is a reason not to do it. For example, if one is playing the game Bullshit, or the game Diplomacy — both of which are sometimes said to be designed to involve lying, or at least to not discourage it. [Schroeder, 2011a, 331]

> This second form of defeat, or something very close to it, is discussed also in the literature on practical reasoning, where it is considered as part of the general topic of "exclusionary" reasons, first introduced by Joseph Raz [...]. Raz provides a number of examples to motivate the concept, but we consider here only the representative case of Colin, who must decide whether to send his son to a private school. We are to imagine that there are various reasons pro and con. On one hand, the school will provide an excellent education for Colin's son, as well as an opportunity to meet a more varied group of friends; on the other hand, the tuition is high, and Colin is concerned that a decision to send his own son to a private school might serve to undermine support for public education more generally.
>
> However, Raz asks us to imagine also that, in addition to these ordinary reasons pro and con, Colin has promised his wife that, in all decisions regarding the education of his son, he will consider only those reasons that bear directly on his

> son's interests. And this promise, Raz believes, cannot properly be viewed as just another one of the ordinary reasons for sending his son to the private school, like the fact that the school provides a good education. It must be viewed, instead, as a reason of an entirely different sort — a "second-order" reason for excluding from consideration all those ordinary, or "first-order" reasons that do not bear on the interests of Colin's son. [Horty, 2007, 14–15]
>
> Beginning with the practical domain, imagine that I have borrowed a book from you. In most situations, the fact that I have borrowed a book from you would give me a reason to return it to you. But suppose I discover that the book I borrowed is one you had previously stolen from the library. In that context, according to Dancy, the fact that I borrowed the book from you no longer functions as a reason to return it to you; in fact, I no longer have any reason to return it to you at all. (ibid.: 20)

The conclusion that we are invited to draw from this is that not only is there no principled account of how various factors contribute to determining what we ought to do, there is no principled account of what the factors are that contribute. What we ought to do is too situationally flexible to be usefully modelled by any once-and-for-all theory.

Instead, the best one can say is that the wise agent sees the situation for what it is and can appreciate the relevant force of the considerations in that context:

> In this respect the judgement as to the rightness of a particular act is just like the judgement as to the beauty of a particular natural object or work of art. A poem is, for instance, in respect of certain qualities beautiful and in respect of certain others not beautiful; and our judgement as to the degree of beauty it possesses on the whole is never reached by logical reasoning from the apprehension of its particular beauties or particular defects. Both in this and in the moral case we have more or less probable opinions which are not logically justified conclusions from the general principles that are recognized as self-evident. [Horty, 2007, 20]
>
> So what we are trying to do is to establish what reasons are present in the case before us. The ability to do this is a sophisticated one, which children develop as they grow

> up; presumably it is one for which some form of training is virtually essential. If we want to know what it is like to have that ability, we could start by asking what it is that competent judges bring to a new case. [...] The particularist will say here that our skills in reason-discernment are not rule-based, meaning by this that we do not extract rules for the operation of reason-giving features from the cases we have come across and then try to subsume new cases under those rules. [...]
>
> [...] Particularists conceive of the knowledge brought to a new case as much more like knowledge-how than like knowledge-that. That is, it is a skill of discernment, not knowledge of a set of true general propositions discovered by thinking about previous cases and applied somehow to new ones. [...] The competent judge is not the person in command of general truths about the behaviour of reasons, all extracted from experience. She is a person who can tell a difference when she comes across it. [Dancy, 2004, 142–3]

So based on reflection on simple examples like this, particularist argue for the conclusion that morality is uncodifiable or shapeless.

But one of the interesting development in recent years is a number of formal systems that can be used to model these examples that motivate particularism. Here I will focus on presenting the ideas of John Horty as they appear in his *Reasons as Defaults* as this theory, in my view, has had the most influence [Horty, 2012]. As the title of Horty's book suggests, his theory makes use of the notion of a reason. Reasons are considerations that count in favor or against some action. Horty's system is designed to allow that the strength of reasons can vary from case-to-case. And it is designed to even allow that what is a reason in one case can fail to be a reason at all in another case. But these changes in strengths of reasons and what is a reason follow from precise principles given by the system. Let's see how this works.

3.2 The Formal Theory

Horty uses of a formal system known as default logic.[40] Default logics were developed originally to understand inferences like the one where we conclude 'Tweety flies' from 'Birds fly' and 'Tweety is a bird'. But Horty

[40] See [Reiter, 1980] for the seminal presentation. See [Makinson, 2005]: ch. 4 for an approachable (albeit slightly idiosyncratic) contemporary introduction.

proposes to use this formalism to model reasons and what we ought to do.

To begin, we have the notion of a default which we can write as '$A \rightarrow B$' to mean that once A has been established one may conclude by default B. We will call A the *premise* of the default and B the *conclusion* of the default. Though this formalism looks like a conditional in ordinary logic, it is not to be understood this way. Instead, it will be used as a kind of principle that encodes which reasons we accept. So for example suppose we think that if John were to promise Mary to help her move that would be a reason for John to help Mary move. We can encode this as the default: John promised to help Mary move $\rightarrow$ John helps Mary move. So if we have the information that John did promise to help Mary move, we can conclude by default 'John helps Mary move'. And in Horty's development of these ideas, what we ought to do is what we can conclude by default. So the theory tells us that given the default and the information that John promised Mary, John ought to help Mary move.

We develop these ideas more systematically by assuming that we have set of sentences, $\mathcal{W}$, that is used to represent the information that we accept and have a set of defaults, $\mathcal{D}$, that represents the "reasons"-principles that we accept. We also assume there is an ordering on this set of defaults, $<$, that tells us about the strength of the reasons. While there are many potential properties one might think this relation has, we only assume that it is transitive (i.e., if $\delta < \delta'$ and $\delta' < \delta''$, then $\delta < \delta''$ for any $\delta, \delta', \delta'' \in \mathcal{D}$) and irreflexive (i.e., $\delta \not< \delta$ for any $\delta \in \mathcal{D}$). Though for certain purposes it may be natural to assume the ordering is connected (i.e, $\delta < \delta'$ or $\delta' < \delta$ for any $\delta, \delta' \in \mathcal{D}$), we do not assume this holds in the general case. So defaults can be tied or can be incomparable. We collect these items together in an ordered 3-tuple $\langle \mathcal{W}, \mathcal{D}, < \rangle$ and call it a *fixed priority default theory.*

The basic idea will be that given our information, $\mathcal{W}$, and the priority ordering, $<$, among defaults, $\mathcal{D}$, some (often proper) subset of $\mathcal{D}$ will be the ones that tell us what ought to be done. This is because some defaults may not apply to a case and some defaults will conflict with defaults that are stronger them. But we only want to pay attention to the applicable defaults that are not in conflict with stronger applicable defaults.

We can call a subset of our set of defaults a *scenario.* So our task will be to define which scenarios are the ones that contain the defaults that tell us what we ought to do. Once we identify that set, we will be able to determine what we ought to do.

We build up to this by noting three features had by defaults that tell you what you ought to do. First, defaults that tell you what you ought to do in a case need to be actually applicable or as we will call it *triggered*. The intuitive idea is that a triggered default is a default that actually represents a reason. For example, if John has not promised Mary to help her move, John's promise is not a reason to help her. But even in cases where John has not actually made the promise, we accept that if he were to make such a promise, it would be a reasons. And defaults are just these kinds of "reason"-principles.[41] So in this example, we have a default concerning John's promises but it is not triggered because John has not made a promise.

We can formalize this in two steps. First, we introduce a pair of functions $Premise$ and $Conclusion$ that respectively return the premise of a default and the conclusion of a default. So for the default δ^*: John promises to help Mary move $\rightarrow$ John helps Mary move, $Premise(\delta^*)$ = John promises to help Mary move and $Conclusion(\delta^*)$= John helps Mary move.

We lift this definition to sets of defaults in the natural way:

for any scenario $\mathcal{S} \subseteq \mathcal{D}$,

$$Premise(\mathcal{S}) = \{Premise(\delta) \mid \delta \in \mathcal{S}\}$$
$$Conclusion(\mathcal{S}) = \{Conclusion(\delta) \mid \delta \in \mathcal{S}\}$$

Next we use these to define the set of triggered defaults are for an arbitrary scenario based on on a theory:

$$Triggered_{\langle \mathcal{W},\mathcal{D},<\rangle}(\mathcal{S}) = \{\delta \in \mathcal{D} \mid \mathcal{W} \cup Conclusion(\mathcal{S}) \vdash Premise(\delta)\}$$

So for example in situation where it is not part of our hard information that John promised Mary, relevant default would not be triggered but in a situation in which it is part of our hard information, it would be triggered.

More generally, we can say that A *is a reason for* B just in case $A \rightarrow B$ is a triggered default. And interestingly, this conception of triggering also correctly allows for the "chaining of defaults". So for example, in a setting where $\mathcal{W} = \{A\}$ and $\mathcal{D} = \{A \rightarrow B, B \rightarrow C\}$, we have it that $Triggered_{\langle \mathcal{W},\mathcal{D},<\rangle}(\mathcal{D}) = \{A \rightarrow B, B \rightarrow C\}$. Intuitively this is because A triggers the first default and the first default then triggers the second.

[41] This shows defaults are not reasons but "reason"-principles. What is a "reason"-principle? We discuss this issue in detail in §3.3

The next concept to introduce is the concept of a conflicting defaults. This is easy to define:

$$Conflicted_{\langle \mathcal{W},\mathcal{D},<\rangle}(\mathcal{S}) = \{\delta \in \mathcal{D} \mid \mathcal{W} \cup \text{ Conclusion}(\mathcal{S}) \vdash \neg \text{ Conclusion}(\delta)\}$$

In a set of defaults some might conflict with others, $Conflicted_{\langle \mathcal{W},\mathcal{D},<\rangle}$ collects these conflicted default together.

The final concept is the concept of one default defeating another. This is to be understood in terms of two things. First, for one default to defeat another, they must be in competition in the sense that one cannot obey both at the same time. So they must be conflicting defaults. Second, to lose the conflict is for the default to be worse according to our ordering. This suggests that:

$$Defeated_{\langle \mathcal{W},\mathcal{D},<\rangle}(\mathcal{S}) = \{\delta \in \mathcal{D} \mid \text{there is a } \delta' \in \text{ Triggered}_{\langle \mathcal{W},\mathcal{D},<\rangle}(\mathcal{S}) \text{ such that (i) } \delta < \delta' \text{ and (ii) Conclusion}(\delta') \vdash \neg\text{Conclusion}(\delta)\}$$

As it turns out this definition is not fully adequate for cases in which there are collections of default which can chain together to create defeating relations. There are a number of known proposal for how to deal with this issue, but I set these aside for now because our main ideas can be illustrated with out these complication.[42]

We make use of these three concepts to define a new operator *Binding* that takes us to the defaults that are triggered, unconflicted, and undefeated:

$$Binding_{\langle \mathcal{W},\mathcal{D},<\rangle}(\mathcal{S}) = \{\delta \in \mathcal{D} : \delta \in Triggered_{\langle \mathcal{W},\mathcal{D},<\rangle}(\mathcal{S}) \text{ and } \delta \notin Conflicted_{\langle \mathcal{W},\mathcal{D},<\rangle}(\mathcal{S}) \text{ and } \delta \notin Defeated_{\langle \mathcal{W},\mathcal{D},<\rangle}(\mathcal{S})\}$$

We will use the label *stable* for when a particular scenario based on a theory is is one that we need to pay attention to when determining what we ought to do. We can say:

$$\mathcal{S} \text{ is a stable scenario based on } \langle \mathcal{W}, \mathcal{D}, <\rangle \text{ iff } \mathcal{S} = Binding_{\langle \mathcal{W},\mathcal{D},<\rangle}(\mathcal{S})$$

This is a so-called "fixed-point" definition. If $\mathcal{S}$ is stable, the idea is that it is a fixed point of the *Binding* operator. Intuitively if the scenario you accept is stable, you have no reason to kick any defaults out or add any more defaults in. In practice, one makes use of this definition by

[42] See [Horty, 2012]: ch. 8 and the citations therein for discussion.

working through each possible set of defaults and checking to see if all of the defaults in it are triggered, unconflicted, and undefeated.[43]

Once we have identified a stable scenario, $\mathcal{S}$, we say an *extension*, $\mathcal{E}$ is what follows from the conclusion of $\mathcal{S}$ together with our hard information, $\mathcal{E} = \{X : \mathcal{W} \cup Conclusions(\mathcal{S}) \vdash X\}$. If there is just one stable scenario for a given theory, we say that what we ought to do is anything that is in the extension of that scenario. But as it happens, there can be multiple stable scenarios.

Plausibly enough, this happens when there are ties or incompatibilities among conflicting defaults. In such a case, we have at least two options. We can either say that what we ought to do is anything that is in the extension of *some* stable scenario or we can say that what we ought to do is anything that is in the extensions of *all* of the stable scenarios. It turns out these correspond to deontic logics that allow moral conflicts and ones that do not allow moral conflicts.

This is already on interesting application of Horty' theory: it allows us to study and compare different approaches to putative moral dilemmas. But we will pass over this interesting application because it is orthogonal to our discussion of particularism and because it is discussed in detail elsewhere in this handbook ([Goble, 2013]). We will therefore assume for simplicity the no-conflict version of Horty's theory that takes what we ought to do in case of multiple extension to be what is in the extension of every stable scenario.

Let us illustrate Horty's approach by looking at an example.[44] Suppose, then, that you have a choice of whether to send your son to School 1 or School 2. When it comes to the cost, School 1 is favored. But when it comes to the quality of education, School 2 is favored. Further, we may suppose the cost provides a stronger reason than the education. So we have the defaults $Cheap(x) \rightarrow Attend(x)$ and $Education(x) \rightarrow Attend(x)$ that represent the idea that a given school being cheap is a reason to attend it and a given school providing high quality education is a reason to attend that school. To simplify things, we instantiate these variables so we have the following defaults:

$$\delta_1 : Cheap(s_1) \rightarrow Attend(s_1),$$

[43]This is not quite Horty's official view. He instead adopts a slightly more complicated notion of a "proper" scenario, see [Horty, 2012, §A.1]. There are also certain more general complications related to the possibility of there being no stable scenarios, see [Horty, 2012, §1.3.2] for a preliminary presentation and [Delgrande *et al.*, 1994] and [Antonelli, 1999] for further discussion.

[44]This example and its development is, of course, inspired by the similar case in [Raz, 2002 1975].

$\delta_2 : Cheap(s_2) \rightarrow Attend(s_2)$,
$\delta_3 : Education(s_1) \rightarrow Attend(s_1)$,
$\delta_4 : Education(s_2) \rightarrow Attend(s_2)$

where s_1 and s_2 are constants for School 1 and School 2 respectively. and we collect these defaults together so that $\mathcal{D} = \{\delta_1, \delta_2, \delta_3, \delta_4\}$. We also have the information that cost is more weighty than education so we have:

$$Education(s_1) \rightarrow Attend(s_1) \sim Education(s_2) \rightarrow Attend(s_2) < Cheap(s_1) \rightarrow Attend(s_1) \sim Cheap(s_2) \rightarrow Attend(s_2)$$

where we understand $\delta \sim \delta'$ to indicate δ and δ' have equal priority.[45]

Finally, our background information includes $Education(s_2)$ and and $Cheap(s_1)$ as well as implicitly the idea that going to one school precludes going to the other. It is easy to see $Education(s_1) \rightarrow Attend(s_1)$ and $Cheap(s_2) \rightarrow Attend(s_2)$ are not triggered. And it is also easy to see that $Education(s_2) \rightarrow Attend(s_2)$ is defeated by $Cheap(s_1) \rightarrow Attend(s_1)$. So the unique proper scenario includes only the default $Cheap(s_1) \rightarrow Attend(s_1)$. This give us an extension that is the closure of $\{Education(s_2), Cheap(s_1), Attend(s_1), \neg Attend(s_2)\}$ and so the result is that you ought send your son to attend School 1 rather than School 2.

This illustrates how the theory can handle conflicting reasons and allow for reasons to be stronger than one another. It can also allow that something can be a reason in one case but not in another simply because it doesn't obtain in the other case. But we have not yet seen how the strengths of reasons can vary from one case to another. And we have not seen how a consideration can obtain in one case and provide a reason while the same consideration obtains in a distinct case but provides no reason. To accommodate these more complex dynamics, we need to introduce some further ideas from Horty's theory.

Conceptually, the key to accommodating these more complex dynamics is a certain picture of what the variability in the strengths of reasons and whether there are reasons amounts to. The picture is that this variability is explained by still further reasons. So for example, there may be a standing reason to not tell a lie, but in a context of playing

[45]Technically, Horty's system as he develops does not have the resources to distinguish equally priority defaults from incomparable defaults. But a simple generalization of Horty's system which takes a "greater-than-or-equal-to" priority relation as basic and modifies the definitions in the obvious ways would allow for this. I assume this richer framework for simplicity of presentation here.

the game Bullshit it may be that this reason is weaker or non-existent. The idea, then is, is that the fact that you are playing Bullshit provides a reason that *attenuates* the strength of the reason to not lie or simply *undercuts* it.

Formally this is accomplished by removing priorities from the structure of the theory and introducing them into the object language in such a way that we are able to reason about them. So we have a so-called variable priority theory which is simply an ordered pair of $\langle \mathcal{W}, \mathcal{D} \rangle$ of hard information and a set of defaults. But we now assume that we are working with a language that has symbols for an ordering and names for the defaults. We will use $\prec$ as an object language symbol for $<$, $\simeq$ as an object language symbol for $\sim$, and introduce a 'd' with a subscript as a name for each default $\delta \in \mathcal{D}$. We add as a further stipulation that every variable priority theory contains as part of its information in $\mathcal{W}$ axioms stating that $\prec$ is a transitive and irreflexive relation. It is easiest to see how these variable priority theories work by returning to our example and considering how it and other variants of it fare.

So to analyze our example, we introduce names into the object for each of our defaults. For each default, δ_i, we introduce the object language name d_i. Next one introduces a new default to the theory:

$$\delta_5\colon\ \top \rightarrow d_3 \simeq d_4 \prec d_1 \simeq d_2$$

δ_5 says that by default that cost is more important than education.

To determine which scenarios are stable for a variable priority theory is a bit more complex than for a fixed priority theory. First, to check whether a scenario is proper, one considers what claims about the priority ordering are in the extension of the scenario. So for example the scenario $\mathcal{S}_5 = \{\top \rightarrow d_3 \simeq d_4 \prec d_1 \simeq d_2\}$ has as its extension $\mathcal{E}_5$ which is the closure of $\{d_3 \simeq d_4 \prec d_1 \simeq d_2\}$ together with the hard information in $\mathcal{W}$. One then considers a fixed priority theory with $\mathcal{W}$ and $\mathcal{D}$ as before but with the ordering $<_5$ that matches (the object language claims about) the ordering given by $\mathcal{E}_5$. We then check to see whether $\mathcal{S}_5$ is a proper scenario in the old sense of the fixed priority theory, $\langle \mathcal{W}, \mathcal{D}, <_5 \rangle$. That is, we considering whether $Binding_{\langle \mathcal{W}, \mathcal{D}, <_5 \rangle}(\mathcal{S}_5) = \mathcal{S}_5$

Let us work through this for our example. Since δ_2 (i.e., $Cheap(s_2) \rightarrow Attend(s_2)$) and δ_3 (i.e., $Education(s_1) \rightarrow Attend(s_1)$) are not triggered, they cannot be included in any proper scenario. On the other hand, δ_1 (i.e., $Cheap(s_1) \rightarrow Attend(s_1)$), δ_4 (i.e., $Education(s_2) \rightarrow Attend(s_2)$), and d_5 (i.e., $\top \rightarrow d_3 \simeq d_4 \prec d_1 \simeq d_2$) are all triggered. Nonetheless, $\{\delta_1, \delta_4, \delta_5\}$ is not a proper scenario because both δ_1 and δ_4 are conflicted given the hard information that $Attend(s_1) \supset$

$\neg Attend(s_2)$. $\mathcal{S}_{4,5} = \{Education(s_2) \rightarrow Attend(s_2), \top \rightarrow d_3 \simeq d_4 \prec d_1 \simeq d_2\}$ is also interestingly not a proper scenario. To see why, notice the derived priority of this scenario is the following

$$Education(s_1) \rightarrow Attend(s_1) \sim_{4,5} Education(s_2) \rightarrow Attend(s_2) <_{4,5} Cheap(s_1) \rightarrow Attend(s_1) \sim_{4,5} Cheap(s_2) \rightarrow Attend(s_2)$$

and the fixed priority theory based on it is $\langle \mathcal{W}, \mathcal{D}, <_{1,5} \rangle$. In this setting, $\mathcal{S}_{4,5}$ is not a proper scenario because δ_4 is defeated (by δ_1) in this scenario based on the fixed priority theory.

$\mathcal{S}_{1,5} = \{Cheap(s_1) \rightarrow Attend(s_1), \top \rightarrow d_3 \simeq d_4 \prec d_1 \simeq d_2\}$ however is a proper scenario because it is proper scenario the resulting fixed priority theory $\langle \mathcal{W}, \mathcal{D}, <_{1,5} \rangle$. Both defaults are triggered, unconflicted, and undefeated. The scenario does not include the triggered default δ_4 but this is acceptable because δ_4 is defeated. Finally, $\mathcal{S}_{1,5}$ is the unique proper scenario as any of the singleton sets would leave out a triggered, undefeated, and unconflicted default.

The resulting extension then tells us what we ought to do is the same as before. Obviously, in this case, all of the added complexity can seem pointless. But in cases where we want to reason about priorities, this extra complexity is worthwhile.

To illustrate this, let us add to the case that we are discussing. Suppose for example that you have also now promised your partner that in matters involving your child you will give more priority to education than cost. Now we might use P to represent that you made this promise to your partner and add the default:

$$\delta_6\colon\ P \rightarrow d_1 \simeq d_2 \prec d_3 \simeq d_4.$$

We introduce d_6 as the object language name for this default.

If we assume that P is part of our hard information, we now have to consider that all of δ_1, δ_4, δ_5, and δ_6, are triggered. This set of defaults is not a proper scenario because it is now conflicted in two ways. As before, δ_1 and δ_4 are conflicted, but we now also have conflict between δ_5 and δ_6. So we know that our proper scenario will have a most one of each of these. And indeed, as is intuitive, there are exactly two proper scenarios: one consisting of δ_1 and δ_5 as before and another consisting of δ_4 and δ_6. The other pairs are, as we might expect, ruled out. Consider for example $\mathcal{S}_{1,6} = \{\delta_1, \delta_6\}$. This scenario is not proper because δ_1 is defeated (by δ_4) in the fixed priority theory based on the derived ordering of this scenario.

We can elaborate this example still further if, for instance, we assume that we take our more specific promise to take priority over our initial

views about the relative strength of cost and education. To do this, we need only add the following:

$$\delta_7 : \top \rightarrow d_5 \prec d_6.$$

If we add this, we now get as may be expected a unique proper scenario, $\mathcal{S}_{4,6,7} = \{\delta_4, \delta_6, \delta_7\}$. This scenario recommends sending one's child to the school that provides the best education, School 2.

Though we have been brief and quite informal, this, I hope,illustrates how we can model the dynamics of how the strengths of reasons can shift. The idea is that reasons themselves explains why in one context a reason can have one strength and in other context in can have a different strength. And Horty works through a variety of other examples in his book (in greater detail than we can do here) and shows just how flexible this framework is at accommodating the variety of dynamics of reasons that have so impressed particularists.

This does not yet give us a way of modelling cases of undercutting. But there are two promising approaches to this. One approach is to take it that there is some place in the ordering that is a threshold in the sense that anything below that spot in the ordering is not a reason. In this setting, one says a default is triggered when it meets the old condition and is above the threshold. This allows us to preserve the idea that reasons are premises of triggered default.

An alternative approach does not make use of a threshold. Instead, the approach says that reasons are defaults that have a special property. We can introduce a predicate *Out* into our language and we can say that if a default's premise obtains and is *Out*, then it is not a reason but if it lacks this property of being *Out* and the premise obtains it is a reason. We then modify the definition of triggering so that a default is triggered when it meets the old definition and is not *Out*.

It should be easy to see how both approaches allow that a consideration can be a reason in one case but not in another. According to the first approach, this is because we can reason about whether a certain default is above or below the threshold. According to the second approach, this is due to our reasoning about whether a certain default has the property of being *Out*. As it turns out, there are interesting differences between these approaches. We will take a look at this in detail later on (§3.4.2).

But for now we pause to take stock of our basic results are. We have seen that a certain picture about the shapelessness of morality is motivated by the fact that there are a variety of relevant normative

considerations that compete to determine what we ought to do, that the strength of these considerations vary from context to context, and that in certain contexts these considerations can even have no strength at all. These phenomena were used by particularist to motivate the idea that there can be no systematic account of the moral. But the formal theory that we have just described can accommodate each of these features with at least as much precision as typically non-particularistic moral theories. This shows one important contribution formal work in deontic logic makes to ethical theory: it undermines a certain argument for particularism. In the remainder of this section, I will further explore how satisfactory this response to the particularist is. And then I will turn to further issues that are raised by the present framework.

3.3 Limitations of the Response to Particularism

Horty's theory gives a systematic account of the shiftiness of reasons and what we ought to do by using a fixed background set of defaults where this set of defaults encodes information not just about which acts to do but also information about the properties of the defaults themselves. Shifts in what reasons there are and how strong reasons are are explained by shifts in hard information and how these pieces of hard information interact with the fixed background of defaults.

But what are these defaults that form a fixed background against which changes are explained? To start, notice that defaults are not reasons. In Horty's theory, reasons are the premises of triggered defaults. This means (the premise of) a default can fail to be a reason by failing to be triggered and this occurs primarily by the premise failing to hold or by the default itself being undercut.

Horty himself describes defaults as generalizations or defeasible principles (cf. especially [Horty, 2012, 16–7 and 42–3]. But what are these? One interpretation is the following:

> ACTUAL NORMATIVE RELATION: defaults model an actual normative relation between two propositions

So this intepretation claims that if $P \to A$ is a default, then (whether P is true or A is true) there is a normative relation between P and A. And when certain facts about the world obtain (e.g., P), this together with the default explains why P is a reason for A.

This is to be contrasted with a second interpretation:

> MODAL-NORMATIVE MIXED RELATION: defaults model under

> what condition it would be the case that the reason-relation holds between two proposition

This interpretation does not entail that any actual normative relation holds between the two propositions.

To see why accepting MODAL-NORMATIVE MIXED RELATION does not (at least without some further assumptions) require one to accept ACTUAL NORMATIVE RELATION, it perhaps helps to consider two examples by way of analogy. First, consider a theory that tells us under what conditions two people would become married. This theory can be informative and interesting for many reasons, but it need not be a theory according to which there is any marital or otherwise interesting actual relationship between people who would be married under certain non-actual conditions. The theory may simply be describing modal-marital mixed relation that is determined by embedding claims about the conditions sufficient for marriage under modal operators.

Or to take an example closer to the normative case, consider a theory that tells us under what conditions P would be believed by John to be a reason for A. No matter how good this theory is at predicting the facts about what normative relations hold in John's belief worlds under various conditions, it gives us no immediate grounds to think any normative relation actually holds between P and A.

So these two interpretations are distinct and the model is, at least initially, neutral about which interpretation is correct. As such, it is neutral on certain further metaphysical questions that the particularist may be interested in. For the particularism central concern may actually be whether there is any actual normative relation that can be systematically theorized about and that determines what we ought to do.

Why might this be so? If ACTUAL NORMATIVE RELATION is incorrect, then it is natural to think facts about what reasons there would be are not explained by (or wholly grounded in or reducible to) genuine normative relations. Now particularist tend to be non-reductivists who believe that normative relations cannot be explained by (or grounded in or reduced to) anything that we can use descriptive language to to talk about. So such a particularist would see claims about what reasons there might be as unexplained by any principles, moral or descriptive. Instead, defaults are just free floating generalizations about what reasons there would be. This, it seems, corresponds to one sense in which the particularist may be interested in a "principle free" approach to ethics.

If this is right, then while Horty's theory is enough to undermine

a certain argument for particularism, it is not itself incompatible with particularism.

So the particularist may agree that their argument fails because it is possible to explain the dynamics of what we ought to do in the manner Horty does. But she may say that this explanation is in tension with particularist explanation only if we accept ACTUAL NORMATIVE RELATION. And this interpretation, she may argue, is less plausible than MODAL-NORMATIVE MIXED RELATION.

To start to show why MODAL-NORMATIVE MIXED RELATION is more plausible than ACTUAL NORMATIVE RELATION, the particularist may point out that the natural way to talk about the connections between the propositions paired by a non-triggered defaults (in English at least) is simply to embed a reason-claim in a counterfactual or other kind of alethic modal expression (e.g, if I were to promise Mary that I will give her \$100, then there would be a reason for me to give Mary \$100). Indeed, it is hard to think of a pretheoretically available term for the kind of normative relation that the first interpretation claims there to be. So we do not seems to think or talk about a relation of the sort envisioned by ACTUAL NORMATIVE RELATION. Furthermore, MODAL-NORMATIVE MIXED RELATION is simpler in the sense that ACTUAL NORMATIVE RELATION is committed to all the claims made by MODAL-NORMATIVE MIXED RELATION together with the claim that some actual normative relation makes these claims true.

According to this particularist reply, what Horty's theory teaches us is that the particularist's argument was not strictly speaking sound. But to avoid the argument's conclusion, one must posit a normative relation that we seemingly heretofore did not think or talk about that determines what reasons we have and what we ought to do.

That said, others have taken a much less skeptical attitude to the kind of normative relation posited by ACTUAL NORMATIVE RELATION. Thomas Scanlon in recent work, for example, goes in for precisely this view and calls the relation R.[46] He writes:

> the essentially normative content of a statement that $R(p, x, c, a)$ [that R holds between a proposition p, an agent x, a context c, and an act a] is independent of whether p holds. This normative content lies in the claim that, whether p obtains or not, should p hold then it is a reason for someone in c to do a. [Scanlon, 2014, 40–1]

[46] An idea like this is found as early as [Chisholm, 1964]. [Horty, 2012, 42–3] also appears to endorse it.

It is however hard to find any *argument* in Scanlon's book that such a normative relation exists rather than there being a mixed modal-normative relation determined by the embedding of a normative claim under modal expressions.[47] This may be because he simply assumes that there is such a normative relation and is defending this idea from objections.[48] While this is a worthwhile project, I do not think it adds any more support for ACTUAL NORMATIVE RELATION. And it provides no response to the concern that we have no evidence that there is such a relation that we have thought or talked about. Indeed, Scanlon's own gloss on what R is simply modally embeds a claim about reasons without doing anything to show us that R is itself a normative relation. So while positing such a relation may ultimately be worthwhile, we still must face up to this consequence.

Of course, there may be still other interpretations of what defaults are that are possible and problematic for the particularist. For example those who are reductivists may have analyses according to which defaults represent certain relations that are in the reductive base and these relations together with certain further facts determine what reasons we have.

If this right, then Horty's theory helps us make dialectical progress by having us focus on providing an interpretation of what defaults are. Since defaults serve a precisely defined role in the theory, this provides some constraints on what it takes for an interpretation to be admissible. Nonetheless, the theory still may allow for a variety of different interpretations corresponding to various particularist and anti-particularist views. It therefore does not settle the debate; rather, it refocuses where the debate should occur.

3.4 Other Problems and Competing Implementations

Here we catalogue some remaining issues and topics that are not directly related to the debate about a particularism but are important for spelling out the correct model of how reasons explain what we ought to do.

3.4.1 Derivative and Non-Derivative Reasons

Our thought and talk about reasons is rich and we often recognize a variety of reasons that are interestingly related to one another. For

[47] Cf. [Schroeder, 2015b, 196].

[48] The best explanation of Scanlon's remarks in my view is that he assumes that there are certain normative relations that hold of necessity so that their holding in one possible world suffices to show they hold in all the worlds.

example, if I promised some friends that I will help them move, I not only take myself to have a reason to help my friends move but I also take myself to have a reason to get up at 6am if that is what it takes for me to help my friends move. There is an interesting relation among these two reasons.[49] The reason to get up at 6am is, in some way, derivative of the reason to help my friends move; the reason to get up at 6am depends on the fact that getting up at 6am is means to helping my friends move. An interesting question, then, is whether the reasons one is modelling using tools like Horty's are reasons of both the derivative and non-derivative sort or just reasons of the non-derivative sort or whether this makes any difference.

The first thing to notice is that it does make a difference. For example, suppose I promised John that I would both take him to the Chinese embassy (to pick up his visa) and take him to LAX. In this case (I hereby stipulate), this is a single promise that I make to John to do two things. Suppose further still that John would not be interested in me merely committing to do one of these acts as he needs me to do both acts for either act to be worthwhile to him (there is no point in going to get the visa if he can't make it to his flight; there is no point in showing up for the flight without a visa). Next suppose that I have promised Mary that I will take her to Burbank airport. Finally, suppose that while I can take John to the embassy and LAX, can take John to the embassy and Mary to Burbank airport, and can take John to LAX and Mary to Burbank airport, I cannot do all three things a once. There isn't time for all that driving.

In this example, we get different results in Horty's system about what ought to be done depending on which reasons we include in our default theory. So far, we have mentioned two reasons. The reason to take John to the embassy and LAX and the reason to take Mary to Burbank. If these are the only two reasons we include, Horty's theory (on its non-conflict allowing interpretation) delivers the results that I ought to either take John to the embassy and LAX or take Mary to Burbank.

But it is very natural to take there to not merely be these two reasons. Instead, there are also some derivative reasons in this case. In particular, given that the promise provides a reason to take John to the embassy and LAX, it also provides a reason to take John to the embassy and a reason to take John to LAX. These are after all (constitutive) means

[49]This is true even if, due to the existence of weighty competing reason, I ought to refrain from helping my friends move and I ought to refrain from get up at 6am.

to those ends.[50] If we include these reasons as additional defaults in Horty's theory, we get different results about what we ought to do. In particular, Horty's theory (on its non-conflict allowing interpretation) does not deliver the result that I ought to either take John to the embassy and LAX or take Mary to Burbank. Instead, we only get the result that I ought to take John to the embassy and LAX or Mary to Burbank and John to the embassy or Mary to Burbank and John to LAX. What this means is that if I were to only take Mary to Burbank, the first way of modelling this cases says that I will have done what I ought to have done while the second way of modelling the case says that I will have failed to do what I ought to do.

Which modelling choice is correct? I myself believe the first modelling choice is correct. What matters fundamentally is that I keep my promises. In this case, I cannot keep both. But there is no grounds for thinking I fail to do what I ought to do when I keeps one of the promises but do not partially fulfill the other. On the assumption that I haven't separately promised to do that act which is the partial fulfillment and on the assumption that there are no other reasons to do that act, there is, in my view, nothing amiss with what I do when I only take Mary to Burbank.[51]

If this is right, this tell us that one must only include the non-derivative reasons when modeling a case using Horty's theory. And this teaches us that there will need to be a substantial role for moral theory in making use of Horty's theory for we need to know what the non-derivative reasons are. This topic is a contested one in moral theory and as such, correctly implementing Horty's theory will be controversial as well.

That said, there may very well be local contexts in which the theory can be used without taking a stand on these issues. And it may be that Horty's theory is a useful tool for adjudicating certain debates about which reasons are derivative and non-derivative because it allows us to see exactly what these different views in moral theory predict about

[50] If one thinks only causal means are supported by derivative reasons, it does no harm to change the example so that we discuss such means throughout

[51] A slightly more complicated version of this objection is needed for the conflict allowing version of Horty's theory: Suppose one promises to do $A \wedge B$ and promised to do $C \wedge D$ where A and C are not compossible but the remaining acts are compossible. Here a default theory that only includes defaults corresponding to non-derivative reasons tells us that $\neg O(B \wedge D)$. On the other hand, a theory that includes defaults for both derivative and non-derivative reasons gets the result that $O(B \wedge D)$. I believe the results of the theory that only includes defaults corresponding to non-derivative reasons is correct.

what we ought to do.

To take one example that illustrates this second point, there is a debate about the reasons that are provided by making promises. One views is that one has a standing reason with regard to promise keeping and that by making a promise to someone in particular, one thereby derivatively acquires a reason to keep the promise to that person. Another view is that fundamentally one has no reason with regard to promise keeping prior to making promises.[52] Rather making a promise brings a (non-derivative) reason into existence. These views differ about whether there is any reason prior to promise making and differ about whether the reason one has because on makes a promise is derivative or non-derivative. As such, these views will end up making different predictions in different cases and Horty's theory provides a precise set of constraints that will allow us to investigate these differences.[53]

3.4.2 Undercutting Defeat and Downward Closure

Let's return now to discuss two different ways of modelling undercutting defeat. According to one way of doing things, a reason is undercut when it is below a threshold in the ordering. According to the other, a reason is undercut when it is has a certain property of being *Out*. So according to the first proposal whether a default is triggered (and in particular whether is not undercut) depends on its place in the ordering. But according to the second proposal whether a default is triggered (and in particular whether it is not undercut) does not depend on its place in the ordering. Much like a default that is not triggered because its premise doesn't hold, a default that is not triggered because it is undercut can occur anywhere in the ordering on the second proposal.

This difference is interesting because the ordering has certain structural properties (transitivity and irreflexivity) that make it so anything that is lower in the ordering than an undercut reason is itself undercut if the first proposal is true. Horty calls this feature the *downward closure* of undercutting. If the second proposal is true, undercutting need not be downwardly closed. So a key question then is "Is undercutting downwardly closed?". According to the first proposal, the answer is 'yes'. According to the second proposal, the answer is 'no'.

[52] Extremes of these different approaches are typical consequentialist accounts that fall in the first camp and the so-called normative powers approach that fall in the second. But this choice point in the theory of promise keeping also applies to other theories.

[53] Cf. [Schroeder, 2007]: ch. 3's discussion of the so-called standard model of normative explanations.

The first answer is supported by what Mark Schroeder calls *The Undercutting Hypothesis* which says "complete undercuttters are simply a limiting case of such partial undercutters (attenuators)" ([Schroeder, 2011a]: 335). The idea here is that partial undercutters or attentuators are things which lower the place of a reason in an ordering. And complete undercutting is just a case where the reason has been lowered to a spot that is sufficiently low. Now one version of this view takes this to be the very bottom of the ordering. According to this view, it would be trivially true that anything lower in the ordering is undercut. But as it happens, there is good reason to think that the threshold in the ordering is not fixed but is instead context-sensitive.[54] In fact, this is a crucial part of Schroeder's view and indeed part of what motivates The Undercutting Hypothesis for him.

He gives the following example in support of this:

> In the basic case, you are standing outside the library, when you see Tom Grabit exit, pull a book from under his shirt, cackle gleefully, and scurry off. This gives you pretty good reason to believe that Tom just stole a book from the library. Case 2 is just the same as the first case, except that Tom has an identical twin, Tim, from whom you can't visually distinguish him. In this case, it has seemed to the judgment of many philosophers that your visual evidence is not a reason to believe that Tom stole a book. Cases like these have been used in order to introduce the notion of undercutting defeat [. . .]
>
> A simple argument, however, strongly suggests that things are more complicated in the Tom Grabit case. Consider a third version of the case, exactly like the other two except that in the third case, in addition to Tim, Tom has a third identical sibling, Tam, from whom you can't visually distinguish him. This third case underwrites a compelling argument against the intuitive judgment that in the second case, your visual evidence was no reason to believe that Tom stole the book. For if you go on to conclude, in the third case, that Tom stole the book, then you are doing worse than if you had gone on to conclude this in the second case. Your reason to believe that Tom stole the book therefore doesn't seem to have gone away in the second case; it merely seems

[54] [Schroeder, 2007]: 5.3 presents his pragmatic approach. See [Snedegar, 2013] for criticism and a contrastive alternative.

> to have gotten substantially weaker. It seems to have been, in Dancy's [...] phrase, attenuated. [...]
>
> And partial undercutting clearly comes in degrees. If the case in which Tom has two identical siblings shows that in the case in which he only has one, you still have a reason to believe that Tom stole a book, then a fourth case, in which Tom has three identical siblings, will show by analogous reasoning that in the two-sibling case, you still have a reason to believe that he stole a book. And if that is right, then we can construct an indefinite chain of increasingly powerful attenuators, each of which will leave you with a reason to believe that Tom stole a book — simply by arbitrarily increasing the size of his sibling cohort. But once we see that a reason can be arbitrarily attenuated, it is natural to contemplate [...] the *undercutting hypothesis* [Schroeder, 2011a, 334–335]

Thus, while complete undercutting really is being at the bottom of the order, on Schroeder's view, the undercutting that we typically detect and talk about will be context-sensitive and depended on the importance of our reasons and what other reasons we think are relevant to deliberation. This supports the claim that undercutting should be a spot in the order because attenuating, it agreed all around, involves changing places in the order.

Against this conception, Horty has argued that there are cases where a reason is undercut but a reason lower than it in the ordering is not. Here is Horty's example:

> Consider a normative interpretation in which a soldier, Corporal O'Reilly, is subject to the commands of three officers. We now take A as a command by the Captain that O'Reilly is to perform some action, where P stands for the proposition that O'Reilly performs that action, so that δ_1 represents the fact that the Captain's command favors P; we take B as a command by the Major that the O'Reilly is not to perform that action, so that δ_2 represents the fact that the Major's command favors $\neg P$; and we take C as a command by the Colonel that O'Reilly is to disregard the Major's command — perhaps the Colonel knows the that Major is drunk — so that δ_3 represents the fact that the Colonel's command favors the exclusion of δ_2. The priority ordering among defaults now corresponds to the rank, and so the authority, of

> the various officers, with the Major outranking the Captain and the Colonel outranking the Major.
>
> Under this interpretation, it seems clear that the downward closure outcome is incorrect. Again, δ_3 provides a reason for excluding δ_2 — the Colonel has ordered O'Reilly to disregard the Major's command; this command cannot, therefore, be taken as a reason for $\neg P$. But it is hard to see why δ_1 should be excluded, or why the Captain's command should be ignored. Imagine O'Reilly trying to explain to the Captain why he has ignored the Captain's command. O'Reilly might say: "The Colonel commanded me to ignore the Major." The Captain could reply: "But I am not the Major." O'Reilly might persist: "The Major outranks you. If I am not supposed to obey even a higher-ranking officer like the Major, why should I obey a lower-ranking officer like you?" But the Captain could again reply: "You were not commanded to ignore orders from the Major and also from all officers of lower rank. That would have been a different command from the one you were actually given, which was simply to ignore orders from the Major." At this point I think the Captain has won the dispute. [Horty, 2012, 133]

What should we make of this counterexample? For my part, I believe that Horty is right about what one ought to do in this case. And that the Captain's reply is convincing. Prima facie, then, we have a counterexample to the first view. The idea is that the Major's reason in higher in the ordering than the Captains reason but it is undercut and the Captains isn't.

I am, however, less than certain about this last claim. The captain is correct when she says "You were not commanded to ignore orders from the Major and also from all officers of lower rank". Indeed, one may be tempted to say that the Colonel said not to pay attention to Major and so thereby treat the Majors commands less seriously than the commands of other officers. If that is true, we could claim that actually the Colonel's command tells us to put the Major lower in the ordering than officers of lower rank and indeed lower than the contextually relevant threshold as well. That is, while we originally start with an ordering according to rank, the ordering changes and then the Major is moved below the threshold. If this is implicit in the case, then the case is no threat to the Undercutting Hypothesis or the downward closure of undercutting.

Against this, one may simply wonder why the case must be inter-

preted in this way. To this, I have no firm reply other than it is difficult to clearly provide a case where only one of these interpretations is admissible. So it is not clear that the Undercutting Hypothesis is false or that undercutting fails to be downwardly closed.

That said, Horty for his part has a reply to Schroeder's case and accordingly the argument for the undercutting hypothesis.. His view is that there may be local contexts in which one does have downward closure. In those context, we should explicitly encodes this in the default information. His idea is that in cases like Schroeder's, our initial ordering comes from reliability considerations which support downward closure but in domains like the domain of military commands such downward closure is not supported.

For my part, I do not think it is clear which view is correct. It is worthwhile for moral theorists to consider what stance they would take on these issues and consider the different predictions that result. If there turn out to be promising theories that require one treatment rather than the other, this would perhaps would be the best evidence for one treatment over the other.

3.4.3 The Accrual of Reasons

We have seen that default theories allow us to answer as well as pose many interesting questions in moral philosophy. As such, they are among the most promising formal tools for exploring ethical issues. Unfortunately, these theories fair poorly with accommodating one rather simple and central class of cases: Sometimes one can have two reasons to do an act and one reason against it. And it can turn out that one ought to do the first act even though each reason to do it is individually worse than the reason against it. The two reasons together, what is often called the *accrual* of these reasons, provides sufficient support to make it so you ought to do the act.

This kind of case, which I think you will agree is mundane, cannot be easily accommodated by default theories.[55] To illustrate the phenomenon, consider the following example that I have given elsewhere:

> Suppose for example that there is a movie theater and a restaurant across town. And suppose that in order to get to that side of town I must cross a bridge that has a $25 toll. The toll is a reason not to cross the bridge. The movie

[55] There is a small literature on this topic in default logic and argumentation theory: [Delgrande and Schaub, 2004; Gómez Lucero *et al.*, 2009; Gómez Lucero *et al.*, 2013; Modgil and Bench-Capon, 2010; Prakken, 2005], and [Verheij, 1995].

> is a reason to cross the bridge and the restaurant is also a reason to cross the bridge. It may be that if there were just the movie to see, it wouldn't be worth it to pay the toll and if there were just the restaurant, it wouldn't be worth it to pay the toll. But given that there is both the movie and the restaurant, it is worth it to pay the toll. [Nair, 2016, 56]

But in default theories each individual reason to cross the bridge will be defeated by the reason not to cross the bridge. So these theories get the result that you ought no to pay the toll.

Now as it happens not all cases of multiple reasons work like this. Here is an example adapted from an early case due to Henry Prakken [Prakken, 2005, §3.1] and cf. [Horty, 2012, 61]:

> Suppose I am deliberating about an afternoon run, and that both heat and rain, taken individually, function as reasons to not run; still, the combination of heat and rain together might function as a weaker reason to not run (say, because the heat is less onerous when there is rain) [Nair, 2016, 59]

And indeed one can easily modify my original example to illustrate this point as well. Simply suppose that there is one seating for dinner and one showing of the movie and they are at the same time so one cannot attend both. In this variant of the bridge case, the two reasons combined provide no additional support for crossing the bridge.

Perhaps, default theories can model these cases if they assume there is always some extra default in cases where having multiple reasons matters. So for example according to this view in the first bridge case one has an extra reason provided by the movie and the restaurant together. But in the second bridge case one does not. This treatment however is inadequate.

While it may be true that two reasons aren't always better than one, it is not true that in cases where two reasons are better than one, this is because there is some further reason that floats free of the original ones. Instead, whether there is a further reason and how strong it is in a given case appears to have a clear explanation. The explanation in the bridge cases has something do with whether one can do attend both dinner and the movie. The explanation in the running cases has something to do with how onerous it is to run in various weather conditions.

Now these are not full explanations and the second one does not seem to have much prospect of being generalized to other cases. But they are nonetheless enough for us to be confident that there is some kind of explanation of what is happening.

The hard theoretical project is, however, providing a plausible and fully general explanation. To date, there are no especially good proposals that fit with default theories. Indeed, there are some grounds for thinking that to model this phenomena one may need to introduce the kinds of quantitative tools familiar from decision theory such as utility function and a probability function. This is because, in my view, the most satisfactory accounts to date make use of such resources.[56] That said, the topic is one that is wide open as of now. It is one that neither work in ethics nor in deontic logic satisfactorily treat. There is substantial room for collaboration.

3.4.4 Other Important Theories

I have focused on Horty's theory here. This is because it is perhaps presented in the most accessible single work. But there are other theories that have many of the advantages of Horty's theory. These theories differ however from Horty's in a variety of ways. Though the brevity with which I will introduce these theories does not do them justice, it is worth at least mentioning them so that the interested reader may look at them in greater detail.

Horty's theory comes from the default logic tradition. As it turns out, this is a rich tradition with a variety of alternatives to Horty's approach. Perhaps, the approach that admits of the easiest comparison is Jörg Hansen's approach which generalizes the approach of Gerhard Brewka and others.[57] Though the details of the formal theories are too complex to describe here, we can look at an important class of cases in which these approaches disagree.

Here is Horty's presentation of a particular version of this class of cases (sometimes called the Order Puzzle) together with his favored verdict about it:

> Once again, we suppose that the agent is the hapless Corporal O'Reilly, and that he is subject to the commands of three superior officers: a Captain, a Major, and a Colonel. The Captain, who does not like to be cold, issues a standing order that, during the winter, the heat should be turned

[56] See [Nozick, 1968] for a prescient discussion of some of these issues. See [Sher, 2019] for an approach that makes use of both probabilities and utilities. [Nair, 2020b] develops an approach that only makes use of probabilities.

[57] See especially [Hansen, 2008]. Brewka's work is developed in the context of modeling agents reasoning about what to believe, see [Brewka, 1994] and [Brewka and Eiter, 2000]. Other important alternative approaches in default logic include [Delgrande *et al.*, 1994; Baader and Hollunder, 1995; Prakken, 2010].

> on. The Major, who is concerned about energy conservation, issues an order that, during the winter, the window should not be opened. And the Colonel, who does not like to be too warm and does not care about energy conservation, issues an order that, whenever the heat is on, the window should be opened. [...]
>
> [...] O'Reilly's job is to obey his orders exactly as they have been issued. If he fails to obey an order issued by an officer without an excuse, he will be court-martialed. And, let us suppose, there is only one excuse for failing to obey such an order: that, under the circumstances, he is prevented from obeying the order issued by this officer by having chosen to obey another order or set of orders issued by officers of equal or higher rank. [...]
>
> Given the set of commands that O'Reilly has been issued in the Order Puzzle, can he, then, avoid court martial? Yes, he can, by [...] obeying the orders issued by the Captain and the Colonel [...] In this scenario, O'Reilly fails to obey the Major's order [...] but he has an excuse: he was prevented from doing so by obeying an order issued by the Colonel, an officer of higher rank. (5) [Horty, 2012, 204–]

Hansen, and other however disagree with this treatment of the case. Here is Hansen's presentation of a seemingly structurally identical case that suggests Horty's approach is incorrect:

> Suppose that if I am attacked by a man, I must fight him (to defend my life, my family etc.). Furthermore, suppose I have pacifist ideals which include that I must not fight the man. Now you tell me to provoke him, which in the given situation means that he will attack me. Let self-defense rank higher than my ideals, which in turn rank higher than your request. Should I do as you request? By the reasoning advocated by Horty, there is nothing wrong with it: I satisfy your request, defend myself as I must, and though I violate my ideals, I can point out to myself that the requirement to fight back took priority. But I think if I really do follow your advice, I would feel bad. I think this would not just be some irrational regret for having to violate, as I must, my ideals, but true guilt for having been tempted into doing something I should not have done, namely provoking the man: it caused the

> situation that made me violate my ideals. [Hansen, 2008, 26]

Both cases are compelling and appear to be a structurally identical. It is an interesting question, then, which approach to choose and why. These very kinds of structures can predictably arise in moral theories that allow of a plurality of reasons that can conflict in complex ways. As such, moral philosophers should also be interested in understanding these kinds of structures and contributing to resolving the question of which way (if any) is best for handling them.

In addition to work in the default logic tradition, there is work in the tradition of argumentation theory that is importantly related to the framework discussed here. Within the formal tradition, [Dung, 1995] presents the classic approach. Within the philosophical tradition, a framework with some similarities to this is developed in the work of John Pollock [Pollock, 1995].[58] The argumentation approach need not conflict with the default approach but it invites slightly different interpretations and can be developed in a conflicting way. For example, Pollock believed that an undercutting defeater must be stronger than the reason it undercuts in order to defeat it [Pollock, 1995, 103–4]. This is not true in Horty's system as undercutting defeaters and the reasons they undercut do not conflict in his system and so their strengths are not especially relevant.[59] But the family of theories itself is very general and can be used to study a variety of perspectives on the issues that we have been discussing.

Finally, there is work in the input/output tradition which has recently been shown to be useful for modeling similar phenomena to the one that Horty's system models. These theories are interesting in their own right as they are formally quite different from other theories and invite different interpretation (see §5.2). Recently, those who make use of this formalism have shown that it can be useful to model contrary-to-duty obligation in a way Horty's system cannot [Parent, 2011]. And they have shown how to offer a distinctive take on priorities and exclusion among reasons that is relevant both to the Order Puzzle and the dispute about whether undercutting defeaters must be stronger than the reasons they undercut [Tucker, 2018].

Each of these perspectives is worth further exploration and engagement from moral philosophers. They often give different verdicts on

[58] See [Dung, 1995]: 4.2 and [Prakken and Horty, 2011] for comparisons between the argumentation approach and Pollock's theory.

[59] See [Horty, 2012, §5.3.2] for discussion

concrete examples of significance. And they often invite different interpretations that may fit better with certain ethical theories. To date, there has been little engagement with these alternatives to Horty's theory by moral philosophers.

4 Individual and Group Obligations

We turn now to a our final main topic. In moral philosophy as well as in deontic logic, it is typical to focus on what a given individual is obligated to do. But there are groups of individuals (or, if you prefer, some individuals) who together act and, seemingly, can be obligated to do various things. What inferential relations, if any, are there between claims about group obligation and claims about individual obligations?

The question is not an idle curiosity. Issues of great moment often can be characterized as involving group behavior and perhaps group obligation and we often take this to tell us something about what individuals are obligated to do.

Felix Pinkert in a recent article provides a helpful case that illustrates some of the issues that are at stake:

> Ann and Ben are owners of two factories which are located opposite each other on a river. Both agents opt for a production process which releases waste chemicals into the river and thereby kill all the fish in the river and destroy the livelihood of a fishing community downstream. The waste from one factory alone would suffice to kill all the fish, and adding the waste from the other factory does no additional damage whatsoever [...]. If Ann or Ben were to unilaterally produce cleanly, this would make their production uncompetitive compared to the other factory, put them out of business, and destroy the livelihood of their employees. However, if they both were to produce cleanly, then this problem would not arise, and both factories would remain in business and the fishing community would flourish. Ann and Ben each employ 100 workers, the fishing community counts 100 people, and all that matters morally in this case are the livelihoods of the workers and fishermen. Further, the only available actions are either to pollute or to produce cleanly. In particular, Ann and Ben cannot come together and suggest and discuss a common strategy.

> [. . .] The Two Factories becomes a challenge for Act Consequentialism only once we assume that Ann and Ben are both "uncooperative", that is, each would pollute even if the other produced cleanly. [. . .]In The Two Factories, it is only if both agents are uncooperative that neither could have improved matters by acting differently and that Act Consequentialism judges that both act rightly. Lastly, Ann and Ben are fully aware of this situation. [Pinkert, 2015, 973–4]

We can ask a number of questions about this example. Are Ann and Ben together obligated to not pollute? If so, how does this affect what Ann is obligated to and what Ben is obligated to do? Moreover how do facts about what Ann will do affect facts about what Ben is obligated to do and vice-versa?

As Pinkert suggests, standard act consequentialism appears to give certain answers to these questions. Standard consequentialism suggest that what we together *can* do is of no special interest to what I am obligated to do. All that matters is what others *would* do given what I do. I must then consider for each act, what others would do if I were to do that act and how good that situation would be. In such a setting, Ann ought to pollute and Ben ought to pollute given that the other in fact will pollute. When it comes to the question of what we together should do, standard consequentialism if it applies to collections of people suggests that that we together ought to not pollute because if we together were to not pollute this would lead to the best outcome.

There are many other versions of cases like this. Some of which do not rely on the idea that others are uncooperative, but instead rely on the idea that no single act makes a difference. Though there are important differences between these kinds of examples, we will not dwell on this here.[60]

Moral philosophers have also suggested certain high level theoretical principles concerning the relationship between what individuals ought to do and what they accomplish by collectively doing what they ought. These principles have comes to be discussed under the banner of *The Principle of Moral Harmony* due to a famous paper by Fred Feldman introducing these ideas and exploring their importance:

> With a few exceptions, moral philosophers seem to be agreed that, at the level of the individual, morality doesn't necessar-

[60] [Regan, 1980] is a seminal discussion of these issues in the context of consequentialist theories. Important recent discussions include [Woodard, 2008; Kagan, 2011; Nefsky, 2012], and [Dietz, 2016].

> ily pay. Hardly anyone who thinks about it seriously would maintain that doing what he morally ought to do invariably benefits the agent more than would some worse alternative. However, when we rise from the level of the individual to the level of the social group, we find that the reverse is true. Quite a few moral philosophers seem to believe that when all the members of a social group do what they morally ought to do, the group as a whole does benefit more than it would have from the performance of any worse alternative set of actions. I shall say that any such view is a version of the Principle of Moral Harmony. [Feldman, 1980, 166-167]

Different moral theories give different answers to the question of whether there is some correct version of the *Principle of Moral Harmony*.[61]

Standard consequentialism may claim that the *Principle of Moral Harmony* is false. In Pinkert's case, act consequentialism suggests each of agent ought to pollute. So if each does what she is obligated to do, they end up polluting. But this is worse than if each failed to do what each was obligated to do; namely, pollute. That said, some care is required in the formulation and evaluation of the *Principle of Moral Harmony*. For as Donald Regan (in [Regan, 1980]) points out, if both agents in fact do acts that result in the best outcome (i.e., both do not pollute), consequentialism says they both acted rightly. So it is not obvious whether consequentialism is incompatible with whatever the precisely formulated and true version of the *Principle of Moral Harmony* is.

We now turn to introducing some theories from deontic logic that bear on these matters.

4.1 Quantified Deontic Logic and Group Agency

There are at least two traditions for thinking about claims about what an individual agent ought to do. According to the simpler of the two, the claim that John ought to do x is to be understood as the claim that it ought to be that John does x. In this setting, we can represent the idea that the collection of John and Bill are obligated to do an act as the claim that it ought to be that John and Bill do this act.

[61] An important early discussion of this principle is [Regan, 1980, especially p. 181ff]. Important recent discussions of moral harmony include [Estlund, 2017; Portmore, 2018; Portmore, 2019] and [Kierland, 2006].

4.1.1 Quantified Deontic Logic

We begin with this perhaps overly simple view of group obligation and see how it relates to our problem. We pay particular attention to the representation of these situations using quantification (e.g., claims like everyone or all of us ought to pollute). Later we turn to more sophisticated approaches to representing agency and obligation of the sort introduced in §2.1.3 and consider how this approaches handles group obligation.

In the factory case, one perspective claims that it is true of each individual (Ann, Ben) that that individual ought to pollute, but it is not true that everyone (Ann and Ben) ought to pollute. According to our scheme for translating these claims in to claims about what ought to be, this means that in this case everyone is such that it ought to be that she pollutes. But it is not the case that it ought to be that everyone pollutes. This suggests that the following claim is false were '$\mathbf{P}x$' is interpreted as 'x pollutes'.

$$\forall x O(\mathbf{P}x) \rightarrow O(\forall x \mathbf{P}x)$$

In a setting in which O is understood to be a modal necessity operator, this claim is an instance of the so-called Barcan formula

This same perspective also suggest that everyone (Ann and Ben) ought to not pollute but it is not true of each individual that she ought to pollute. According to our translation scheme, this means that it ought to be that everyone doesn't pollute. But it is not the case that everyone is such that it ought to be that she doesn't pollute. This suggests then that the following claim is false:

$$O(\forall x \neg \mathbf{P}x) \rightarrow \forall x O(\neg \mathbf{P}x)$$

This claim is an instance of the so-called converse Barcan formula.

Of course, this way of translating things may be incorrect. And we will consider other formalizations and interpretations of the example later. But for now let us focus on this initial first pass.

Compared to propositional deontic logic, quantified deontic logic has received much less attention. So much of the ground that we will cover now is speculative and draws on analogies from other fields.

A standard way of thinking of debates about the Barcan formulae given the usual semantics for alethic modality is to think of it as concerning whether there could've been a larger or smaller number of things. Very roughly, in the context of alethic modality, failures of the Barcan formula (if there are any) are taken to illustrate that there could be more

things than there (actually) are. And failures of the converse Barcan formula (if there are any) are taken to illustrate that there could be fewer thing than there (actually) are. Let us examine this and consider what it might teach us about failures of the Barcan formulae in the deontic context.[62]

It is easiest to appreciate putative counterexamples to the Barcan formula and its converse if we work with their equivalent formulations involving existential quantifiers and possibility modals. In this setting the Barcan formula looks like this:

$$\Diamond\exists x(\alpha) \rightarrow \exists x\Diamond(\alpha)$$

where α is an arbitrary (possibly open) formula. To see why some reject, this claim, consider that it is possible for Wittgenstein to have had a daughter (though actually he did not). Where we let '$\mathbf{D}xy$' stand for x is y's (biological) daughter and we let w name Wittgenstein, we have it that $\Diamond\exists x\mathbf{D}xw$. On the other hand, Wittgenstein doesn't have a daughter. Is there, nonetheless, someone or something that is possibly his daughter? It seems 'no'. Who or what would it be? Certainly not any of Wittgenstein's actual children or anyone else's children. Plausibly, then $\neg\exists x\Diamond\mathbf{D}xw$.

How does this example fare in the deontic setting? It certainly seems permissible for Wittgenstein to have a (biological) daughter. And it may seem plausible then that there is among the ideal worlds one in which Wittgenstein has a daughter. But is there someone who exists here and now who is permitted to be Wittgenstein (biological) daughter? This question is harder to answer, but plausibly the answer is 'no' if we accept the counterexample to the alethic version of the Barcan formula. After all, in that case we think $\neg\exists x\Diamond Dxw$. But this means there is nothing that is such that it could be Wittgenstein's daughter. And plausible a generalization of the idea that 'ought' implies 'can' suggests that no one is permitted to be a way they cannot be.

Let us consider how this line of thought fares in our pollution example. There it is permissible that someone refrain from polluting (because it is permissible that Ann and Ben together refrain from polluting). But

[62]The discussion below ignores the considerable resources of available to necessitists who accept the Barcan formulae and believe that there is a necessary framework of objects. I myself am sympathetic to this approach and believe many of these issues may be helpfully explored within a necessitist friendly deontic logic. But I do not explore this here for the sake of introducing our topic in a way that is connected to failures of simpler, more familiar formulae. See [Williamson, 2015] for a systematic defense and development of the necessitist picture. The examples below are inspired by Williamson's discussion.

it is not true of anyone that it is permissible for that person to refrain from polluting (because Ann ought to pollute and Ben ought to pollute). But the model from the alethic case does not help us to understand why these claims hold. It is not as though there is an deontically ideal world where one of Ann or Ben exists even though one of them does not exist in the actual world. The existence or non-existence of certain agents is simply not a relevant feature of these kinds of collective action problems.

That said, while existence and non-existence may be of little relevance to the cases of collective action that we have in mind, there are interesting issues in ethics concerning the existence and nonexistence of agents. And it may be that the present model is better suited to exploring those issues. We will briefly consider this in §4.3. But for now, let us continue to explore how failures of the converse Barcan formula in the alethic setting may be relevant to our topic.

So consider a standard putative counterexample to the converse Barcan formula. Again, it helps to consider the version of it that makes use of existential quantifiers and possibility modals:

$$\exists x \Diamond(\alpha) \rightarrow \Diamond \exists x(\alpha)$$

Consider now that someone is such that it is possible that she does not exist. That seems true of me, for example. We might formalize this as follows $\exists x \Diamond(\forall y(x \neq y))$. On the other hand, it is impossible for there to be something that is not identical to anything. So $\neg\Diamond\exists x(\forall y(x \neq y))$.

Let us consider a similar case in a deontic context. Plausibly, there is someone who is permitted not to exist. Perhaps, this is someone whose life has been filled with nothing but pain, is incapable of forming interesting relationships, etc. Whatever the exact details are, it is highly plausible there are creatures that do not have lives worth living and plausibly it is true of them that they are permitted to not exist. So $\exists x P(\forall y(x \neq y))$ where 'P' is here interpretted as the operator 'it is permissible that'. And if we accept the counterexample to the alethic Barcan formula, it also suggests we have a counterexample to the deontic one. This is because we generally have as a theorem $P(\alpha) \rightarrow \Diamond(\alpha)$. So since $\neg\Diamond\exists x(\forall y(x \neq y))$, $\neg P\exists x(\forall y(x \neq y)))$.

Yet once again, it is not clear how this kind of understanding of failures of the Barcan formula help us to understand the target case of interest. There we have it that someone (each of Ann and Ben) is such that they are permitted to pollute. But we want to reject the claim that it is permissible for a person to pollute (because it is required that Ann and Ben (together) not pollute). What we just saw is that this can happen when one of the people we are talking about in the actual world

fails to exist in an ideal world. But this, as before, is not relevant to the pollution case. Ann and Ben, we may assume, exist in the actual world as well as all ideal worlds.

One reaction to this is that this shows what is wrong with the view that the group ought to not pollute but each individual ought to pollute. But this reaction is mistaken for at least two reasons. First, while the model does not allow for the verdicts the consequentialists gives about the case in any sensible way, the model also does not represent much of the interesting underlying structure of the case. For example, we have no representation of the causal relations or counterfactual relations between the individual and group actions. Second, it may be that this representation of what agents ought to do in terms of what ought to be the case is incorrect.

One way of trying to improve on these shortcomings is to consider what richer theories such as Lou Goble and Sven Ove Hansson's theories (described in §2.1.2) say about this case. But we will not dwell on how these theories treat these cases because their application to these cases is straightforward and can be easily checked by the reader by consulting the more detailed description of the theories provided in §2.1.2.

Instead, I will simply state some relevant facts about how these theories treat the cases that we are discussing: First, both theories allow that it may be that some individuals ought to do something while the group ought to do an incompatible act (roughly, this is because of the way in which these theories fail to validate DEONTIC INHERITANCE). Second, since Goble's theory has a counterfactual structure, it does allow us to model some aspects of the interaction of groups and individuals. Third, Hansson's theory (on the simplest way of modelling these cases) suggests that if a group ought to do something then at least one individual in the group ought to do it as well (roughly, this is because Hansson's theory validates DISJUNCTIVE DIVISION). Fourth, neither theory tells us much about the *Principle of Moral Harmony*. In principle both allow failures of it.

With these results in mind, we turn to a more complex theory that allows for a richer representation of the structure of our cases.

4.1.2 Group Agency

We can also explore these ideas in the context of the agency based deontic logic discussed in §2.1.3. Though we will briefly restate the crucial features of this framework, the reader who has not looked at §2.1.3 will need to consult it for a better understanding of these features and what

motivates them. Our primary focus is on how group obligation and the representation of group agency can be introduced in that framework.

Let us recall, then, the basics of Horty's framework. We have frames with the following structure: $\langle Tree, <, Agent, Choice, Value\rangle$. $Tree$ is a set containing elements that we call moments and use 'm', perhaps together with a subscript, to refer to a moment. $<$ is an ordering on $Tree$ that ensures the moments form a "tree-like" branching time structure. A set of moment that form a complete unbroken path through the tree (more precisely, a maximal linearly ordered set of moments) is called a *history* and we use 'h', perhaps together with a subscript, to refer to a history. $Agent$ is a set of agents and $Choice$ is a function that maps an agent and a moment to a partition of histories through that moments. The cells of the partition are *acts* and we use 'K', perhaps together with a subscript, to refer to an act. Finally, $Value$ assigns numbers to histories where these numbers are understood to be a measure of the value of each history. We use M to designate a model based on such a frame where we assume the usual semantics for our underlying propositional language and present the remaining interest semantic clause below.

Using these resources, we saw that we were able to analyze obligations and actions of individuals. We begin with the following useful definitions:

- $H_m = \{h \mid m \in h\}$
- $|A|^{M,m} = \{h \in H_m \mid A \text{ is true relative to } M, m/h\}$
- $Choice^{m,\alpha}(h) = \{K \in Choice^{m,\alpha} \mid h \in K\}$
- $State^{m,\alpha} = Choice^{m,Agent-\{\alpha\}}$
- For sets of histories, P and Q, $P \leq Q$ iff $Value(h) \leq Value(h')$ for all $h \in P$ and $h' \in Q$
- For acts, K_1 and K_2, $K_1 \preceq K_2$ iff $K_1 \cap S \leq K_2 \cap S$ for each $S \in State^{m,\alpha}$
- $K_1 \prec K_2$ iff $K_1 \preceq K_2$ and $K_2 \npreceq K_1$
- $Optimal^{m,\alpha} = \{K \in Choice^{m,\alpha} \mid \text{there is no } K' \in Choice^{m,\alpha} \text{ such that } K \prec K'\}$

This allows us to define what it takes for α to see to it that A, $[\alpha$ *cstit:* $A]$, and for α to be obligated to do A, $\odot[\alpha$ *cstit:* $A]$:

$[\alpha$ *cstit:* $A]$ is true relative to $M, m/h$ iff $Choice^{m,\alpha}(h) \subseteq |A|^{M,m}$

$\odot[\alpha$ *cstit:* $A]$ is true relative to $M, m/h$ iff $K \subseteq |A|^{M,m}$ for each $K \in Optimal^{m,\alpha}$

This picture can be generalized to apply not just to individual agents, but collections of agents. To do this, we need to generalize some of our definitions slightly so that they apply not to elements of *Agent* (i.e., individual agents) but subsets of *Agent* (i.e., groups of agents). The basic idea is just to take the acts available to the group to be the conjunction of acts available to the individuals. In order for this idea to be sensible, it is assumed that at a given moment the action of one agent does not affect the acts available to another agent at that moment. We call this the independence of agents property and require that models satisfy it. These ideas are made precise as follows:

$Select_m$ is a set of functions s that for each $\alpha \in Agent$, $s(\alpha) \in Choice^{m,\alpha}$

Independence of agents is satisfied iff $\bigcap_{\alpha \in Agent} s(\alpha) \neq \varnothing$ for each moment m and $s \in Select_m$

$Choice^{m,\Gamma} = \{\bigcap_{\alpha \in \Gamma} s(\alpha) \mid s \in Select_m\}$

Once we have the *Choice* function generalized to groups, all of the previous definitions and clauses can be directly applied to the choice sets of groups.

Let us consider then what this theory says about our guiding example and the *Principle of Moral Harmony*. Figure 4 represent the simple two agent case that is discussed by Pinkert in the present framework. In this situation, α faces of choice to do K_1 or K_2 and β faces a choice of K_3 or K_4. The result of α performing K_1 is α polluting which we represent as A_p; analogously β performing K_3 results in β polluting which we represent as B_p. The other acts result in the agent not polluting and instead producing cleanly. According to Pinkert's telling, when $A_p \wedge B_p$, only the 100 workers in both factories do well. This is what occurs in h_2 so its has a value of 200. When $A_p \wedge \neg B_p$, the 100 workers in α's factories do well, but the 100 workers in β's factories do not because they are at a competitive disadvantage because they do not pollute. This is what occurs in h_1 so it has a value of 100. Analogously, when $\neg A_p \wedge B_p$, only the 100 workers in β's factories do well. This is what occurs in h_3 so it a value of 100. Finally, when $\neg A_p \wedge \neg B_p$, the 100 workers in both factories

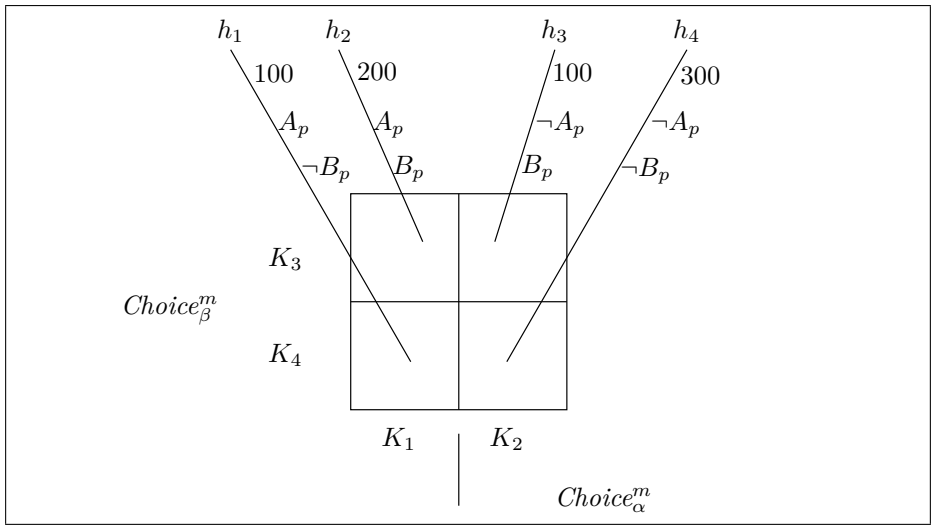


Figure 4: Two Agent Pollution Decision Tree

and the 100 people in the fishing village do well because the waters are not polluted. This is what occurs in h_4 so it has a value of 300.

If we consider now what α and β together ought to do, we must begin with checking what the choice set of this group is. According to our definitions $Choice^{m_1,\{\alpha,\beta\}} = \{K_1 \cap K_3, K_1 \cap K_4, K_2 \cap K_3, K_2 \cap K_4\}$. In this case each of these choice determines a unique history: $K_1 \cap K_3$ results in h_2; $K_1 \cap K_4$ results in h_1; $K_2 \cap K_3$ results in h_3; $K_2 \cap K_4$ results in h_4. So determining the optimal act is just a matter of comparing the values of these histories and it is clear, therefore, that $K_2 \cap K_4$ is the unique optimal act. Thus what α and β together ought to do is not pollute. In other words, $\odot[\{\alpha, \beta\}$ *cstit:* $\neg A_p \wedge \neg B_p]$.

But what should each of these agents do individually? Begin with α who faces a choice between K_1 and K_2. To determine, which act is optimal, we need to consider whether either act is state-wise dominant. In this case, the states are just given by the other agents acts. So our question reduces to whether some act results in the better outcome no matter what the other agent does. And it is easy to see both K_1 and K_2 are non-dominated in this sense. In the state K_3, K_1 produces a better outcome (h_2); in K_4, K_2 produces the better outcome (h_4). Thus both K_1 and K_2 are optimal. So we have it that $\neg \odot [\alpha$ *cstit:* $\neg A_p]$ and $\neg \odot [\alpha$ *cstit:* $A_p]$. It is easy to check analogously that $\neg \odot [\alpha$ *cstit:* $\neg B_p]$ and $\neg \odot [\alpha$ *cstit:* $B_p]$.

We noted that in certain setting consequentialism not only suggested that α and β together ought to not pollute, but also that α ought to

pollute and β ought to pollute. But notice Horty's theory does not have this result in this example. α is merely permitted to pollute and same for β.

The difference between the results that Horty's theory provides and the results that the consequentialist provide is that Horty treats what other agents will do as a genuinely open matter. The theory does not take it as given for the purpose of evaluating what one agent ought to do what the other agent will do. This is reflected in the fact that one must consider the value of one's act in states where the other agent does the action that the case suggest they will not do. Consequentialism however takes as settled what others will do (at least in cases where one cannot affect what other will do).[63]

What of the *Principle of Moral Harmony*? It is hard to say whether or not Horty's system validates this principle without trying to state the principle more precisely. But we can notice one feature of Horty's theory that suggests it may validate something like the *Principle of Moral Harmony*. Suppose a group of individuals can realize the uniquely best history through their actions. In this setting, each individual doing what they ought to do can never ensure that this history fails to obtain.

We can informally see this as follows: Suppose h is the history that is the (uniquely) most valuable. According to Horty's theory, it can never be that if each of a set of individuals does what they ought to do, they will have seen to it that h fails to obtain. This is because any act/state pair compatible with h is guaranteed to be non-dominated. And as such, there will be no unique optimal act that ensures that h fails to hold for each individual. Could it be that though each of a pair of individuals, α, β, has no optimal act that ensures h does not obtain, they each have exactly one optimal act, K_1 and K_2 respectively, which are individually compatible with h but when done in tandem $(K_1 \cap K_2)$ ensures h does not obtain? No, for if K_1 is uniquely optimal for α, this means the act/state pair $K_1 \cap K_2$ is strictly preferred to any alternative act available to α together with K_2. But now notice that K_2 must not be the uniquely optimal act for β. This is because we know that there is a state S such that $K_1 \cap S$ is compatible with h (since K_1 is compatible with h). Given what we have said before, this state includes an act, K_3, that is available to β and distinct from K_2 (because $K_1 \cap K_2$ is incompatible with h). But this means that K_2 cannot be the unique

[63] Horty notices this, shows how to define a consequentialist notion of obligation in his system, and shows that his formalization get the same results as consequentialism. See [Horty, 2001]: §5.4 for the treatment of individual obligation and §6.2-6.3 for group obligation.

state-wise non-dominated act as $K_1 \cap K_3$ is not dominated by $K_1 \cap K_2$. So it must not be that each individual doing what they ought precludes the best history.

This informal demonstration rides roughshod over some important subtleties and complications. But hopefully, it is enough to give a sense of why the result holds in Horty's theory. An important topic for further consideration is the forulation and evaluation of more precise versions of the *Principle of Moral Harmony*. It would be especially worthwhile to consider how an argument due to Donald Regan (1980) that claims to prove that certain natural formulations of the *Principle of Moral Harmony* cannot be true fares once regimented in a system like Horty's. That argument and related discussion since suggest a principled inadequacy with any approach that only represents and evaluates actions (rather than also considering the attitudes of agents).[64]

So Horty's frameworks allow us to study the interaction of individual obligation and group obligation. And it also gives us some insight into whether the *Principle of Moral Harmony* holds. As it turns out, there are other logics in this tradition as well. While we do not have the space to consider them here, studying these logics in greater detail is a natural next step to take in trying to gain greater philosophical and formal traction on the issues raised by collective action problems.

4.2 A Speculative Alternative Framework

We close our discussion of collective action problems by proposing (but not developing) an alternative framework for understanding these problems that is suggested by some remarks from Derek Parfit. Parfit in his unpublished but much cited paper "What We Together Do" writes:

> My suggested version of AC [act consequentialism] may seem incoherent. Suppose that, in Regan's Case, we both do *A*. On my suggestion, though each of us acts rightly, we act wrongly. This may seem impossible. How can truths about each not be true of us?
>
> With some truths, this is not possible, Thus, if each is old, we cannot be young, Youth is a property of individuals: we together cannot be young. But other properties are different. Even though each is weak, we together may be strong.
>
> My suggestion is of this second kind. [Parfit, 1988, 8–9]

[64] See [Portmore, 2019]: §5.3 for discussion.

The case Parfit is discussing is much like our factory case. He is considering the view that it is right for each individual to pollute but it is wrong for us to pollute. Parfit here appears to be drawing our attention to the fact that some predicates like the predicates 'weak' and 'strong' are what linguists call *non-distributive* predicates. They are predicates that can apply to groups of people without applying to their members ('strong' in Parfit's example) or to individuals who make up a collection of people without applying to the collection ('weak' in Parfit's example). Parfit is suggesting 'right' and 'wrong' may exhibit such behavior.

We can take a leaf then from the study of these non-distributive predicates and see how it may apply to studying deontic logic. And it turns out there are two main approaches to this issue. According to one approach, we enrich the kind of things there are to include not just ordinary individuals but also some things for when we are talking about several individuals. For example, perhaps there are also sets of individuals. Other leading proposals have been mereological sums, or events with mereological or set-theoretic structure. Whichever approach one favors, the idea is developed by claiming a predicate may apply to this entity (i..e, this set, mereological sum, etc.) without applying to the ordinary individuals. This approach is often favored by linguists.[65] And, in effect, this is the approach that we have already pursued in the previous subsection when discussing Horty's treatment of group obligation. There we assigned obligations to sets of individuals as distinct from the individuals who make up the set.

But there is another approach to these cases favored in certain circles of philosophical logic.[66] It is the approach that does not posit anything in addition to individuals, but rather involves the idea that we may attribute plurally some features to individuals. When some marbles are scattered, being scattered is a property of these marbles. One does not posit that there is some further entity, the set of marbles or the marble fusion, that has this property. These ideas are developed formally in theories that allow for plural predication and quantification.

My suggestion, inspired by Parfit's comments, is that it would be worthwhile to develop deontic logics with the resources of plural predication and quantification in order to model our reasoning about group obligation in collective action problems of the sort that we have been dis-

[65]Important representatives of this approach include [Link, 2002] and [Schein, 1993]. [Schein, 2006] provides a survey.

[66]Important representatives of this approach include [Boolos, 1984; Boolos, 1985; Yi, 1999; Oliver and Smiley, 2001; Rayo, 2002]. [Rayo, 2007] provides a useful survey and [McKay, 2006] is a rich book length treatment.

cussing. As far as I know, there has been no discussion in the literature of deontic logic about plural predication and quantification.

Here is not the place to make the first steps toward developing such a logic. But in recent years there has been some useful work in analogous fields. In particular, there has been work on understanding the interaction of plural predication and quantification and alethic modal logic. This work can serve as a tentative guide for us. Some of the highlights of this literature include discussion comparing the relation of an individual being one of some individuals and the relation of identity in modal context and include discussion of Barcan formulae involving plural quantifiers.[67]

While quite speculative, my hunch is that this approach will be fruitful for exploring the relationship between individual obligation and the obligations of groups and for exploring the status of the *Principle of Moral Harmony*.

This concludes our discussion of collective action problems. As can be seen, the formal literature on this topic is not as rich as the literature we have discussed about our previous topics. This is especially true of approaches that are related to quantification in deontic logic. The field is wide open.

4.3 The Ethics of Existence

We close our discussion by turning away from issues related to collective action problems and instead returning to consider whether there are any other examples from ethics that not only lead to failures of the deontic Barcan formulae but also do so in a way that makes sense given the traditional interpretation of what these formulae say. Since these formulae on their standard interpretation tell us something about what there is or what exists, a good place to look is at issues concerning existence.

And in moral philosophy, there are a number important issues related to existence and non-existence. Perhaps most famous among them is the so-called non-identity problem discussed originally in Derek Parfit's *Reasons and Persons* ([Parfit, 1986]: ch. 16). Though we do not have the space here to explore this topic in any detail, we will briefly present the problem and connect it to our discussion of deontic Barcan formulae.[68]

[67] See [Bricker, 1989], [Linnebo, 2016], [Uzquiano, 2004], [Uzquiano, 2011],and [Williamson, 2010].

[68] A recent survey is [Roberts, 2015]. But see also [Gardner, 2015] and the citations therein.

Molly Gardner in a recent article provides a nice explanation of the non-identity problem:

> Consider the following two cases:
>
> > *Case 1.* During her pregnancy, Alice takes a drug that she knows will cause her child, whom she names Alex, to develop poor health. Despite his poor health, Alex has a life worth living. He would have had a higher level of well-being if Alice had not taken the drug.
> >
> > *Case 2.* Barbara uses *in vitro* fertilization and screens the embryos for a particular gene that causes poor health. When she finds an embryo with that gene, she implants it and discards the rest. The selected embryo becomes a child named Billy, who develops poor health. Having poor health causes Billy to experience exactly the same hardships, pain, and suffering that having poor health causes Alex to experience. However, like Alex, Billy has a life worth living.
>
> Intuitively, both Alice's action and Barbara's action are objectionable. The objection to Alice's action is that she has clearly harmed her child. Since Barbara's action is similar — it affects Billy in almost the same way that Alice's action affects Alex — we might be tempted to think that the objection to Barbara's action is also grounded in harm.
>
> Nevertheless, there is a difference between Case 1 and Case 2. The difference is that, although Alex would still have existed had his poor health *not* been induced, Billy is *non-identical* to anyone who would have existed, had his poor health not been selected for. [...]
>
> Many philosophers argue that this metaphysical difference makes a moral difference. According to the *counterfactually worse-off condition* on harming, an action harms someone only if it makes her worse off in at least some respect than she would have been, had the action not been performed. Alice's action satisfies this condition [...] However, Barbara's action does not satisfy this condition. Billy's life is worth living, and plausibly, having a life worth living is not worse

> for Billy in any respect than not existing; therefore, Billy is no worse off in any respect than he would have been, had Barbara not selected for poor health. [. . .]
>
> But if Barbara's action does not harm Billy, then we seem to be at a loss to justify the intuition that, in much the same way that Alice's action is objectionable, Barbara's action is also objectionable. [...] The problem of either accounting for this appearance that the individual was wronged or explaining it away is the *non-identity problem.* [Gardner, 2015, 428–429]

Crucial to the case is the interaction between a certain metaphysical claim — the claim that in w, $\neg\exists x\ x =$ Billy where w is the world that would result if Barbara had chosen a different embryo — and a moral principle that depends on it, the *counterfactually worse off condition* on harming. As Gardner says, the combination of these claims leads to the suggestion that Barbara's action is not wrong or at least seemingly that the grounds for it being wrong is much different than the grounds for Alice's action being wrong.[69]

In the actual world, there is Billy and Billy has poor health. Evidently many believe that it ought to be that Barbara choose a different embryo much like how it ought to be that Alice not take the drug. So it ought to be that there is someone who is both Barbara's child and not Billy and (given the set up of the case) there is no one who is Billy. That is, $O(\exists x\ x$ is Barbara's child $\wedge\ x \neq$ Billy $\wedge\ \neg\exists y\ y =$Billy). This is equivalent to the claim that $O(\exists x\forall y\ (x$ is Barbara's child $\wedge\ x \neq$ Billy $\wedge\ y \neq$ Billy)) and the claim that $O(\forall y\exists x\ (x$ is Barbara's child $\wedge\ x \neq$ Billy $\wedge\ y \neq$ Billy)).

From $O(\exists x\forall y(x$ is Barbara's child $\wedge\ x \neq$ Billy $\wedge\ y \neq$ Bill)), the fact that obligations entail permissions, and the Barcan formula, we have $\exists xP(\forall y(x$ is Barbara's child $\wedge\ x \neq$ Billy $\wedge\ y \neq$ Bill)). If we suppose Barbara's only actual child is Billy, then there is no one who is possibly Barbaras's child and not Billy.[70] So it seems $\neg\exists x\Diamond(\forall y\ (x$ is Barbara's child $\wedge\ x \neq$ Billy $\wedge\ y \neq$ Billy)) and it is therefore plausible that $\neg\exists xP(\forall y$ (x is Barbara's child $\wedge\ x \neq$ Billy $\wedge\ y \neq$ Billy)). Thus, reasoning about

[69] Of course, consequentialist will say the reason why both acts are wrong is they do not bring about the best outcome and so they will deny that what makes Alice's act wrong is that it causes harm. Gardner, on the other hand, rejects the *counterfactually worse-off condition* on harming. We do not discuss the details of these approaches here.

[70] This is true at least in the contingentist setting we adopted for the purposes of exposition in this article.

this example appears to suggest that the Barcan formula fails. And it fails because of the fact that in some deontically ideal world there is something that exists that does not exist in the actual world.

Similar results hold for the converse Barcan formula. From $O(\forall y \exists x (x \neq \text{Billy} \wedge y \neq \text{Billy}))$ and the converse Barcan formula, we have it that $\forall y O(\exists x(x \neq \text{Billy} \wedge y \neq \text{Billy}))$. But consider Billy as value for y. Given that Billy cannot fail to be self-identical, the embedded formula cannot be true in any world (ideal or otherwise). Thus, reasoning about this cases appears to suggest the converse Barcan formula fails. And it fails because there is something the exists in the actual world that fails to exist in all deontically ideal worlds.

Unlike the examples involving collective action problems, this example of a failure of the deontic Barcan formulae does seem to be sensibly related to the standard interpretation of the underlying formalism. What is ethically relevant is the existence of something in ideal worlds that doesn't exist in the actual world (Barcan formula) and the nonexistence of something in ideal worlds that does exist in the actual world (converse Barcan formula)

This shows that there are examples of significance in ethics that provide motivation for exploring deontic logics where Barcan formulae fail.[71] But can deontic logic provide any insight to help resolve or evaluate important arguments in moral philosophy concerning the ethics of existence? Here it is harder to say because, as far as I know, the topic has not been systematically explored to date.

What's more, there are a number of other important problems in moral philosophy that concern the existence and nonexistence of individuals.[72] For example, there is the paradox of mere addition and the related repugnant conclusion that concerns issues about whether to bring more people into existence who have lives worth living.[73] All of these topics have, to my knowledge, received little to no systematics study by deontic logicans

[71] Or the exploration of necessitist deontic logics that allow for an alternative approach to putative failures of the Barcan formulae. We have, to repeat, focused on raising these issues in a contingentist manner only to simplify exposition.

[72] These issue primarily originate from [Parfit, 1986]: Part 4's discussion of population ethics but [Naverson, 1967] is an earlier work that explores some issues related to existence. [Greaves, 2017] is a recent survey of views from the perspective of formal axiology.

[73] The paradox originates in [Parfit, 1986, ch. 17-19]. [Arrhenius *et al.*, 2017] provides a recent survey.

5 Further Topics

I have chosen to look in detail at three particular issues in moral philosophy and consider their connection to various frameworks in deontic logic. But there are many other topics where there have been fruitful interaction or their could be fruitful interactions between deontic logicians and moral philosophers. We briefly consider a small selection of disparate topics.

5.1 The Logical Form of Obligation

In ethics and deontic logic, there has been considerable study of the logical form of claims about obligation. One issue that has been of interest in both fields for some time is how best to represent what an agent ought to do as opposed to what ought to be the case. Another issue that has received attention in ethics in the last decade or two is about the proper scope of the obligation operator with respect to conditionals in statements of what rationality requires of us. Let us briefly look at each of these topics.

We speak of things that ought to be or occur (e.g. it ought to be that there is world peace) and we also speak of what agents ought to do (e.g., we ought to keep our promises). How are these related? On one hand, it seems as though when we discussion what an agent ought to do we are interested in a certain relation between an agent and an action. But since there appears to be no relation at all between an agent an action involved in claims about what ought to be, claims about what ought to be and what we ought to do, it seems, have substantially different logical forms. On the other hand, we might try to reduce what we ought to do to what ought to be. And a popular and tempting analysis (which we have used in various places in this article) claims that what we ought to do is just a special case of what ought to be; we ought to do something exactly when it ought to be that we do it. And there are many possible views in between these two extremes.

A variety of evidence from logic, semantic, and ethics have been brought to bear on choosing among these options.[74] And while the dominant tradition in logic and semantics has been to adopt the tempt-

[74]Some important discussions from the linguistic, logical, and philosophical traditions include [Broome, 1999; Broome, 2013; Castañeda, 1981; Chisholm, 1964; Chrisman, 2012; Finlay, 2014; Finlay and Snedegar, 2014; Geach, 1982; Horty, 2001; Ross, 2010; Schroeder, 2011b; von Wright, 1951; Wedgwood, 2006], and [Williams, 1981 1980]

ing analysis, a variety of formal approaches have been developed that depart from it including the framework discussed in §2.1.3 of this article.

Finally, a number of important issues in ethics may turn on which view about these issues is correct. In a recent article Mark Schroeder [Schroeder, 2011b] mentions at least four: the viability of certain meta-ethical analyses, the viability deontology, the adequacy of the agent-neutral/agent-relative reasons distinction, and the prospects of wide-scope accounts of rationality.

This last item on Schroeder's list — the prospects of wide-scope accounts of rationality — is a topic that has received much discussion in its own right.[75] The issue here concerns what is rationally required of us. To focus on just one kind of example, we know that when someone believes that they ought to intend to do something and fails to intend to do it, something has gone wrong with them. They are, in some way, irrational. The following claim, if true, would explain why this agent is irrational:

> If S believes that S ought to intend to do x, then S is rationally required to intend to do x

But many have objected to this claim on the grounds that it involves a kind of illegitimate bootstrapping. Suppose S is not rationally required to intend to do x and indeed S ought not to intend to do x. But nonetheless, suppose S now comes to believe that S ought to intend to do x. The above principle then says that S now in fact is rationally required to intend to do x. But S should not, now, intend to do x. Instead, S ought to stop falsely believing that S ought to intend to do x.

In light of this, some have proposed the following alternative account of our rational requirements:

> It is rationally required that if S believes that S ought to intend to do x, then S intend to do x

Here, the thought goes, one can respond to this requirement by dropping the belief (as S should in some cases like the one described in the previous paragraph) or by forming the intention (which will be appropriate in many other cases).

The debate concerning this requirement as well as some related requirements turns not just on the correct verdict about certain examples

[75] A small sampling of important papers on this topic include [Broome, 1999; Kolodny, 2005; Kolodny, 2007; Schroeder, 2009]. [Kiesewetter, 2017] and [Lord, 2018b] are recent book length exploration of these and related topics.

but also how conditionals and requirements interact in the face of further factual and normative information. Arguments for and against the above view often make some assumptions about under what conditions one can conclude that one is rationally required to intend to do x given a proposed requirement and given certain pieces of factual and normative information.

This of course is an issue that has been studied in detail by deontic logicians. And work in ethics has benefited from this work in deontic logic even if there are still some insights from logic that have yet to be noticed.[76]

5.2 Input/Output Logic

In a series of important papers in the early 2000s, David Makinson and Leendert van der Torre initiated the study of what they called "Input/Output Logics" [Makinson and van der Torre, 2000; Makinson and van der Torre, 2001]. Since then, these logics have been used to study permissions, contrary-to-duty obligation, and a variety of other topic of importance in deontic logic and ethics. This handbook provides a detailed discussion of these logics [Parent and van der Torre, 2013].

Here we consider whether there are some additional applications for this formal theory. In particular, from the perspective of the moral philosopher what are some interpretation of the formalism of input/output logic that could allow us to use this powerful formal theory to understand issues in ethics.

In their original paper, van der Torre and Makinson introduce the study of input/output logic as follows:

> Imagine a black box into which we may feed propositions as input, and that also produces propositions as output. Of course, classical consequence may itself be seen in this way, but it is a very special case, with additional features — inputs are also themselves outputs, since any proposition classically implies itself, and the operation is in a certain sense reversible, since contraposition is valid. However, there are many examples without those features. Roughly speaking, they are of two main kinds.
>
> The box may stop some inputs, while letting others through, perhaps in modified form. Inputs may record reports of

[76] Some examples of work that appeals to certain logical principle include [Broome, 2007; Broome, 2013; Schroeder, 2009; Schroeder, 2015a; Lord, 2018b], and [Lord, 2018a]. There are many others.

> agents, of the kind 'according to source i, x is true', while the box may give as output either x itself, a qualified version of x, or nothing at all, according to the identity of i. Or it might give output x only when at least two distinct sources vouch for it, and so on. [...] In these examples, the outputs express some kind of belief or expectation.
>
> Again, inputs may be conditions, with outputs expressing what is deemed desirable in those conditions. The desiderata may be obligations of a normative system, ideals, goals, intentions or preferences. In general, a fact entertained as a condition may itself be far from desirable, so that inputs are not always outputs; and as is widely recognised, contraposition is inappropriate for conditional goals.
>
> Our purpose is to develop a general theory of propositional input/output operations, covering both kinds of example. [Makinson and van der Torre, 2000, 383–4]

Input/output logic aims to be a general theory of how one can transform some input to get some output where, it appears, that the transformations that we aim to model are those involved in inference.

In the article in this handbook dedicated to input/output logic and its relevance to deontic logic, Xavier Parent and Leendert van der Torre tell us "the first objective states that detachment is viewed as the core mechanism of the semantics of normative reasoning" where by detachment they mean the way in which one reaches conclusions about what one is obligated to do from conditional claims about what one is obligated to do together with other information ([Parent and van der Torre, 2013]: 502). They also tell us:

> The view of logic underpinning the I/O framework is very different. Its role is not to create or determine a distinguished set of norms, but rather to prepare information before it goes in as input to such a normative code, to unpack output as it emerges and, if needed, coordinate the two in certain ways. A set of conditional norms is, thus, seen as a transformation device, and the task of logic is to act as its "secretarial assistant". [Parent and van der Torre, 2013, 506]

The idea then is that for an arbitrary set of conditional norms, input/output logic helps us prepare inputs to the set of norms and retrieve outputs about what we ought to do given these inputs.

From the perspective of a moral philosopher, it is not immediately clear how best to understand these claims in a way that allows us to see what light input/output logic sheds on issue in moral philosophy. But two interpretations stand out corresponding to the two comments by van der Torre and Parent.

The first comment from Parent and van der Torre and the comment from Makinson and van der Torre highlight certain aspects of agents' reasoning about what they are obligated to do. As such, we can take them to be modelling good forms of reasoning about obligation. This in itself is not a model of what we in fact ought to do.

The second comment from Parent and van der Torre, on the other hand, more naturally suggests that the theory models what we in fact ought to do given that a certain set of conditional norms is in force (and certain other factual or normative information holds). This in itself is not a model of good forms of reasoning about obligation.

These different interpretations should be kept conceptually separate as one is a model of an epistemological issues while the other is a model of a metaphysical issue. And it may be that different desiderata bear on the adequacy of the model depending on whether it is a model of good forms of reasoning or a model of what is obligatory.

That said, we should not make too much of the difference either. One may take a model of good forms of reasoning about obligation to give us insight into what we in fact ought to do if we are willing to accept certain theories of obligation. For example, according to some, facts about what we ought to do are explained by facts about *correct reasoning* about what (we ought) to do.[77] If we accept such a view, then a model of good forms of reasoning about obligation also turns out to be a model of what we are in fact obligated do to do.

One may take it that the metaphysical model gives us insight into good forms of reasoning about obligation if we apply it to the set of conditional norms that the agent we are modeling accepts (as opposed to the conditional norms that obtain) and the set of inputs that the agent has received (as opposed to the facts).

Though we do not need to decide which of these interpretations is correct in order to see the interest in input/output logic, having a settled interpretations may be helpful: It may make it easier for those in ethics to see what lessons they can draw from the analysis of various deontic phenomena given by input/output logic. It may suggest further ways in which this formalism could be used to advance our understanding of

[77]See [Williams, 1981 1979; Setiya, 2014], and [Way, 2017]

issues in ethics.[78][79]

5.3 Contrary-to-Duty Obligation

The logic of contrary-to-duty obligation (obligations conditional on failing to do what one is obligated to do) has been extensively studied in deontic logic. However the topic has largely been neglected by moral philosophy. This may be because moral philosophers believe that moral theory need not give an account of contrary-to-duty obligations: The point of moral theory, some may believe, is to tell us what we are (unconditionally) obligated to do in a given situation. Once we fail to do what we are obligated to do, there may be new things that we ought to do, but there is no reason for an ethical theory to say now what we ought to do given that we fail to do something that we ought to do.[80]

As plausible as this perspective might seem, it does not reflect the richness of human concern. We not only go about trying to do the right thing, but we also go about developing contingencies plans about what to do in case we fail to do the right thing: An alcoholic on the road to recovery will of course know that she ought not to drink tonight and plan not to do so. But a truly wise alcoholic on the road to recovery will also know that she ought to call her sponsor if she does drink and put in place a plan to do so conditional on her drinking. It is therefore a worthwhile area for philosophical investigation.

In recent years, there has been a number of papers that have explored more specific issues in moral theory with an eye to the special

[78] van der Torre also suggests to me that [Liao *et al.*,] and [Benzmüller *et al.*,] are important contributions to ethical theorizing that make use of the input/output framework. The fact that discussion of these approaches from computer science to reasoning about applied problems are not included here, highlights the very specific perspective on ethical theory that I have adopted.

[79] Similar issues of interpretation arise for the so-called "Theory of Joining Systems" [Lindahl and Odelstad, 2013] but this theory has not been applied extensively to issues in ethics and its implications for ethics are less well understood (at least by the author of this paper).

[80] Alternatively, moral theorists may believe that once they have given a theory of what we are obligated to do in an arbitrary case, this theory can be trivially generalized to also be a theory of contrary-to-duty obligation. While this may work for certain moral theories, the task is not trivial for reasons akin to the ones mentioned in §2.1.1 and §2.3.2. Notably, political philosophers often discuss "non-ideal theory" ([Valentini, 2012] provides a survey) and take it to be an important and non-trivial subject matter concerning what we ought to do given that we have failed to live up to the demands of "ideal theory" (which tells us, e.g., what the perfectly just state consists in). It is curious, then, that moral philosophers have taken considerably less interest in this subject matter ([Korsgaard, 1986] is an important exception).

features of contrary-to-duty obligations. The actualist/possibilist debate [Kiesewetter, 2015; Kiesewetter, 2018; White, 2017], rational requirements [Comesaña, 2015], and moral particularism [Parent, 2011] have all been the subject of some recent work connecting these issues to contrary-to-duty requirements. But the area is worth much deeper study by moral philosophers.

5.4 Moral Conflicts

This handbook devotes an entire article to cases of moral conflict [Goble, 2013]. That rich article introduces a variety of important formal systems and compares the forms of reasoning and argument that they sanction. I recommend it to anyone interested in the topic of moral conflicts.

One area however that is under theorized in the formal tradition is the relationship between certain kinds of (reactive) attitudes such as guilt, resentment, indignation, etc. and obligation. While thinking about appropriate attitudes may not be relevant for many topics that deontic logians are interested in, it is relevant to the topic of moral conflicts: A leading idea from Bernard Williams and Ruth Barcan Marcus has been that it is appropriate to have these reactive attitudes in cases of moral conflict no matter what one does. They have taken this as evidence that these cases involve genuinely conflicting all-things-considered obligations.[81]

What lies behind this idea is a commitment to a principled connection between obligations and reactive attitudes. So a worthwhile project for further work is the development of a logic that can model appropriate reactive attitudes and their connection to obligation.[82] This will allow us to see the assumptions that are made by Williams/Marcus-style arguments. And it will allows us to explore what the logic of appropriate reactive attitudes is and how it relates to the logic of obligations. We may then compare the different predictions made by those who accept or reject the existence of moral dilemmas.

[81] This argument is sometimes known as the argument from "moral residue(s)". See [Williams, 1988 1965] and [Marcus, 1980].

[82] A notable exception to the trend of ignoring attitudes in deontic logic is Paul McNamara's work. McNamara explores the logic of what he calls aretaic attitudes (e.g., praise and blame) in the context of understanding supererogatory action. See [McNamara, 2011a] for the formal theory and [McNamara, 2011b] for further exploration of the philosophical importance of this theory.

6 Conclusion

Though deontic logicians and moral philosophers have not interacted at the depth that we might hope and aspire to in the future, there have been a number of fruitful interactions. We have chosen to focus primarily on three topics of interest in moral philosophy and a few representative theories in deontic logic that address these issues. But we have also seen that these topics and theories by no means exhaust the interesting terrain. The field is ripe for further productive interactions that will help us to better understand our reasoning about what is right and wrong and better develop our theories of what is right and wrong.

References

[Antonelli, 1999] G. Aldo Antonelli. A directly cautious theory of defeasible consequence for default logic via the notion of a general extension. *Artificial Intelligence*, 109:71–109, 1999.

[Arrhenius *et al.*, 2017] Gustaf Arrhenius, Jesper Ryberg, and Torbjörn Tännsjö. The repugnant conclusion. *Stanford Encyclopedia of Philosophy* editor Edward Zalta spring 2017 `https://plato.stanford.edu/archives/spr2017/entries/repugnant-conclusion/`

[Baader and Hollunder, 1995] Franz Baader and Bernhard Hollunder. Priorities on defaults with prerequisites, and their applications in treating specificity in terminological default logic. *Journal of Automated Reasoning*, 15:41–68, 1995.

[Benzmüller *et al.*,] Christoph Benzmüller, Xavier Parent, and Leendert van der Torre. Designing normative theories of ethical reasoning: Formal framework, methodology, and tool support. *CoRR*.

[Boolos, 1984] George Boolos. To be is to be a value of a variable (or to be some values of some variables). *Journal of Philosophy*, 81:430–449, 1984.

[Boolos, 1985] George Boolos. Nomalist platonism. *Philosophical Review*, 94:327–44, 1985.

[Bradley, 2006] Ben Bradley. Against satisficing consequentialism. *Utilitas*, 18:97–108, 2006.

[Braham and van Hees, 2015] Matthew Braham and Martin van Hees. The formula of universal law: A reconstruction. *Erkenntnis*, 80:243–260, 2015.

[Brewka and Eiter, 2000] Gerhard Brewka and Thomas Eiter. Prioritizing default logic. In Steffen Hölldobler, editor, *Intellectics and Computational Logic: Papers in Honor of Wolfgang Bibel*, pages 27–45. Kluwer Academic Publishers, 2000.

[Brewka, 1994] Gerhard Brewka. Reasoning about priorities in default logic. In *Proceedings of the Twelveth National Conference on Artificial Intelligence*

(AAAI-94), page 940–945, 1994.

[Bricker, 1989] Philip Bricker. Quantified modal logic and the plural de re. *Midwest Studies in Philosophy*, 14:372–394, 1989.

[Broome, 1999] John Broome. Normative requirements. *Ratio*, 12:398–419, 1999.

[Broome, 2007] John Broome. Wide or narrow scope? *Mind*, 116:359–370, 2007.

[Broome, 2013] John Broome. *Rationality Through Reasoning*. Wiley-Blackwell, Oxford, 2013.

[Brown, 2018] Campbell Brown. Maximalism and the structure of acts. *Noûs*, 52:752–771, 2018.

[Burgess, 2009] John Burgess. *Philosophical Logic*. Princeton University Press, Princeton, 2009.

[Bykvist, 2002] Krister Bykvist. Alternative actions and the spirit of consequentialism. *Philosophical Studies*, 107:45–68, 2002.

[Cariani, 2013] Fabrizio Cariani. 'Ought' and resolution semantics. *Noûs*, 47:534–558, 2013.

[Cariani, 2016a] Fabrizio Cariani. Consequence and contrast in deontic semantics. *Journal of Philosophy*, 113:396–416, 2016.

[Cariani, 2016b] Fabrizio Cariani. Deontic modals and probability: One theory to rule them all. In Nate Charlow and Matthew Chrisman, editors, *Deontic Modality*, pages 11–46. Oxford University Press, Oxford, 2016.

[Carlsson, 1999a] Erik Carlsson. Consequentialism, alternatives, and actualism. *Philosophical Studies*, 96:253–268, 1999.

[Carlsson, 1999b] Erik Carlsson. The oughts and cans of objective consequentialism. *Utilitas*, 11:91–96, 1999.

[Carr, 2015] Jennifer Carr. Subjective *Ought*. *Ergo*, 2:678–710, 2015.

[Castañeda, 1981] Hector Castañeda. The paradoxes of deontic logic. In Risto Hilpinen, editor, *New Studies in Deontic Logic*, pages 37–86. Reidel, Dordrecht, 1981.

[Chang, 1997] Ruth Chang, editor. *Incommensurability, Incomparability, and Practical Reason*. Harvard University Press, Cambridge, 1997.

[Charlow, 2016] Nate Charlow. Decision theory: Yes! truth conditions: No! In Nate Charlow and Matthew Chrisman, editors, *Deontic Modality*, pages 47–81. Oxford University Press, Oxford, 2016.

[Chellas, 1969] Brian Chellas. *The Logical Form of Imperatives*. PhD thesis, Philosophy Department, Stanford University, 1969.

[Chisholm, 1964] Roderick Chisholm. The ethics of requirement. *American Philosophical Quarterly*, 1:147–153, 1964.

[Chrisman, 2012] Matthew Chrisman. 'Ought' and control. *Austrialasian Journal of Philosophy*, 90:433–451, 2012.

[Cohen and Timmerman, 2016] Yishai Cohen and Travis Timmerman. Actualism has control issues. *Journal of Ethics and Social Philosophy*, 10, 2016.

[Comesaña, 2015] Juan Comesaña. Normative requirements and contrary-to-duty obligations. *The Journal of Philosophy*, 112:600–626, 2015.

[Curran, 1995] Angela Curran. Utilitarianism and future mistakes: Another look. *Philosophical Studies*, 78:71–85, 1995.

[Dancy, 2004] Jonathan Dancy. *Ethics without Principles*. Oxford University Press, Oxford, 2004.

[Dancy, 2017] Jonathan Dancy. Moral particularism, 2017.

[Delgrande and Schaub, 2004] James Delgrande and Torsten Schaub. Reasoning with sets of defaults in default logic. *Computational Intelligence*, 20:56–88, 2004.

[Delgrande *et al.*, 1994] James Delgrande, Torsten Schaub, and W. Ken Jackson. Alternative approaches to default logic. *Artificial Intelligence*, 70:167–237, 1994.

[Dietz, 2016] Alex Dietz. What we together ought to do. *Ethics*, 126:955–982, 2016.

[Dung, 1995] Phan Minh Dung. On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming, and n-person games. *Artificial Intelligence*, 77:321–357, 1995.

[Estlund, 2017] David Estlund. Prime justice. In Kevin Vallier and Michael Weber, editors, *Political Utopias*, pages 35–55. Oxford University Press, Oxford, 2017.

[Feldman, 1980] Fred Feldman. The principle of moral harmony. *The Journal of Philosophy*, 77:166–179, 1980.

[Feldman, 1986] Fred Feldman. *Doing the Best We Can: An Essay in Informal Deontic Logic*. D. Reidel Publishing Company, Dordrecht, 1986.

[Finlay and Snedegar, 2014] Stephen Finlay and Justin Snedegar. One ought too many. *Philosophy and Phenomenological Research*, 89:102–124, 2014.

[Finlay, 2014] Stephen Finlay. *Confusion of Tongues*. Oxford University Press, New York, 2014.

[Gardner, 2015] Molly Gardner. A harm-based solution to the non-identity problem. *Ergo*, 2, 2015.

[Geach, 1982] Peter T. Geach. Whatever happened to deontic logic? *Philosophia*, 11:1–12, 1982.

[Goble, 1990a] Lou Goble. A logic of *good*, *should*, and *would* part i. *Journal of Philosophical Logic*, 19:169–199, 1990.

[Goble, 1990b] Lou Goble. A logic of *good*, *should*, and *would* part ii. *Journal of Philosophical Logic*, 19:253–276, 1990.

[Goble, 1996] Lou Goble. Utilitarian deontic logic. *Philosophical Studies*, 82:317–357, 1996.

[Goble, 2013] Lou Goble. Prima facie norms, normative conflicts, and dilemmas. In Dov Gabbay, John Horty, Xavier Parent, Ron van der Meyden, and Leon van der Torre, editors, *Handbook of Deontic Logic and Normative Systems*, volume 1, pages 241–352. College Publications, Milton Keynes, 2013.

[Goldman, 1976] Holly Goldman. Dated rightness and moral imperfection. *Philosophical Review*, 85:449–487, 1976.

[Goldman, 1978] Holly Goldman. Doing the best one can. In Alvin Goldman and Jaegwon Kim, editors, *Values and Morals*, pages 185–214. D. Reidel Publishing Company, Dordrecht, 1978.

[Gómez Lucero *et al.*, 2009] Mauro Gómez Lucero, Carlos Chesñevar, and Guillermo Simari. Modelling argument accrual in possibilistic defeasible logic programming. In *Symbolic and Quantitative Approaches to Reasoning with Uncertainty*, page 131–143, 2009.

[Gómez Lucero *et al.*, 2013] Mauro Gómez Lucero, Carlos Chesñevar, and Guillermo Simari. Modelling argument accrual with possibilistic uncertainty in a logic programming setting. *Information Sciences*, 228:1–25, 2013.

[Greaves, 2017] Hilary Greaves. Population axiology. *Philosophy Compass*, 12:e12442, 2017.

[Greenspan, 1978] Patricia Greenspan. Oughts and determinism: A response to goldman. *Philosophical Review*, 87:77–83, 1978.

[Gustafsson, 2014] Johan Gustafsson. Combinative consequentialism and the problem of act versions. *Philosophical Studies*, 167:585–596, 2014.

[Hansen, 2008] Jörg Hansen. Prioritized conditional imperatives. *Autonomous Agents and Multi-Agent Systems*, 17:11–35, 2008.

[Hansson, 2001] Sven Ove Hansson. *The Structure of Values and Norms*. Cambridge University Press, New York, 2001.

[Hansson, 2013] Sven Ove Hansson. Alternative semantics for deontic logic. In Dov Gabbay, John Horty, Xavier Parent, Ron van der Meyden, and Leon van der Torre, editors, *Handbook of Deontic Logic and Normative Systems*, volume 1, pages 445–498. College Publications, Milton Keynes, 2013.

[Hooker and Little, 2001] Brad Hooker and Margaret Little, editors. *Moral Particularism*. Oxford University Press, Oxford, 2001.

[Horty, 2001] John Horty. *Agency and Deontic Logic*. Oxford University Press, Oxford, 2001.

[Horty, 2007] John Horty. Reasons as defaults. *Philosopher's Imprint*, 7:1–28, 2007.

[Horty, 2012] John Horty. *Reasons as Defaults*. Oxford University Press, Oxford, 2012.

[Hurka, 1990] Thomas Hurka. Two kinds of satisficing. *Philosophical Studies*, 59:107–111, 1990.

[Jackson and Pargetter, 1986] Frank Jackson and Robert Pargetter. Oughts, actions, and actualism. *The Philosophical Review*, 95:233–255, 1986.

[Jackson, 1985] Frank Jackson. On the semantics and logic of obligation. *Mind*, 94:177–195, 1985.

[Jackson, 2014] Frank Jackson. Procrastinate revisited. *Pacific Philosophical Quarterly*, 95:634–647, 2014.

[Jeffrey, 1990] Richard Jeffrey. *Logic of Decision*. Chicago University Press,

Chicago, 2 edition, 1990.

[Jennings, 1974] R.E Jennings. A utilitarian semantics for deontic logic. *Journal of Philosophical Logic*, 3:445–456, 1974.

[Kagan, 2011] Shelly Kagan. Do I make a difference? *Philosophy and Public Affairs*, 39:105–141, 2011.

[Kierland, 2006] Brian Kierland. Cooperation, 'ought morally', and principles of moral harmony. *Philosophical Studies*, 128:381–407, 2006.

[Kiesewetter, 2015] Benjamin Kiesewetter. Instrumental normativity: In defense of the transmission principle. *Ethics*, 125:921–946, 2015.

[Kiesewetter, 2017] Benjamin Kiesewetter. *The Normativity of Rationality.* Oxford University Press, Oxford, 2017.

[Kiesewetter, 2018] Benjamin Kiesewetter. Contrary-to-duty scenarios, deontic dilemmas, and transmission principles. *Ethics*, 129:98–115, 2018.

[Kolodny, 2005] Niko Kolodny. Why be rational? *Mind*, 114:371–385, 2005.

[Kolodny, 2007] Niko Kolodny. How does coherence matter? *Proceedings of the Aristotelian Society*, 107:229–263, 2007.

[Korsgaard, 1986] Christine Korsgaard. The right to lie. *Philosophy and Public Affairs*, 15:325–349, 1986.

[Krantz *et al.*, 2007] David Krantz, R. Duncan Luce, Patrick Suppes, and Amos Tversky. *Foundations of Measurement*, volume 1. Dover Publishing, Mineola, reprint edition, 2007.

[Lewis, 1974] David Lewis. Semantic analyses for dyadic deontic logic. In Sören Stenlund, editor, *Logical Theory and Semantic Analysis*, pages 1–14. D. Reidel, Dordrecht, 1974.

[Lewis, 2001 1973] David Lewis. *Counterfactuals.* Blackwell Publishing, Oxford, 2 edition, 2001 [1973].

[Liao *et al.*,] Beishui Liao, Marija Slavkovik, and Leendert van der Torre. Building jiminy cricket: An architecture for moral agreements among stakeholders. *CoRR.*

[Lindahl and Odelstad, 2013] Lars Lindahl and Jan Odelstad. The theory of joining-systems. In Dov Gabbay, John Horty, Xavier Parent, Ron van der Meyden, and Leon van der Torre, editors, *Handbook of Deontic Logic and Normative Systems*, volume 1, pages 545–634. College Publications, Milton Keynes, 2013.

[Lindner and Bentzen, 2018] Felix Lindner and Martin Mose Bentzen. A formalization of Kant's second formulation of the categorical imperative. In *Deontic Logic and Normative Systems: DEON 2018*, pages 211–225, 2018.

[Link, 2002] Godehard Link. The logical analysis of plurals and mass terms. In Paul Portner and Barbara Partee, editors, *Formal Semantics*, pages 127–146. Blackwell Publishing, Oxford, 2002.

[Linnebo, 2016] Öystein Linnebo. Plurals and modals. *Canadian Journal of Philosophy*, 86:654–676, 2016.

[Lord, 2017] Errol Lord. What you are rationally required to do and what you

ought to do (are the same thing!). *Mind*, 126:1109–1154, 2017.

[Lord, 2018a] Errol Lord. The explanatory problem for cognitivism about practical reason. In Conor McHugh, Jonathan Way, and Daniel Whiting, editors, *Normativity*, pages 137–61. Oxford University Press, Oxford, 2018.

[Lord, 2018b] Errol Lord. *The Importance of Being Rational*. Oxford University Press, New York, 2018.

[MacFarlane, 2000] John MacFarlane. *What Does It Mean to Say Logic Is Formal?* PhD thesis, University of Pittsburgh, 2000.

[Makinson and van der Torre, 2000] David Makinson and Leendert van der Torre. Input/output logics. *Journal of Philosophical Logic*, 29:383–408, 2000.

[Makinson and van der Torre, 2001] David Makinson and Leendert van der Torre. Constraints for input/output logics. *Journal of Philosophical Logic*, 30:155–185, 2001.

[Makinson, 2005] David Makinson. *Bridges from Classical to Nonmonotonic Logic*. College Publications, Milton Keynes, 2005.

[Marcus, 1980] Ruth Barcan Marcus. Moral dilemmas and consistency. *Journal of Philosophy*, 77:121–136, 1980.

[McKay, 2006] Thomas McKay. *Plural Predication*. Oxford University Press, Oxford University, 2006.

[McKeever and Ridge, 2006] Sean McKeever and Michael Ridge. *Principled Ethics*. Oxford University Press, Oxford, 2006.

[McNamara, 2011a] Paul McNamara. Praise, blame, obligation, and dwe. *Journal of Applied Logic*, 9:153–170, 2011.

[McNamara, 2011b] Paul McNamara. Supererogation, inside and out. In Mark Timmons, editor, *Oxford Studies in Normative Ethics*, volume 1, pages 202–235. Oxford University Press, Oxford, 2011.

[Meachem and Weisberg, 2011] Christopher Meachem and Jonathan Weisberg. Representations theorems and the foundations of decision theory. *Australasian Journal of Philosophy*, 89:641–663, 2011.

[Modgil and Bench-Capon, 2010] Sanjay Modgil and Trevor Bench-Capon. Integrating dialectical and accrual modes of argumentation. In *Proceedings of the 2010 Conference on Computational Models of Argument*, page 335–346, 2010.

[Moore, 1962 1903] G.E Moore. *Principia Ethica*. Cambridge University Press, Cambridge, 1962 [1903].

[Nair, 2016] Shyam Nair. How do reasons accrue? In Errol Lord and Barry Maguire, editors, *Weighing Reason*, pages 56–73. Oxford University Press, New York, 2016.

[Nair, 2020a] Shyam Nair. Fault lines in ethical theory. In Douglas Portmore, editor, *The Oxford Handbook of Consequentialism*, pages 67–92. Oxford University Press, New York, 2020.

[Nair, 2020b] Shyam Nair. "Adding Up" Reasons: Lessons for Reductive and

Non-Reductive Approaches. Forthcoming *Ethics*, 2021.

[Naverson, 1967] Jan Naverson. *Morality and Utility*. John Hopkins Press, Baltimore, 1967.

[Nefsky, 2012] Julia Nefsky. Consequentialism and the problem of collective harm. *Philosophy and Public Affairs*, 39:364–395, 2012.

[Nolt, 2018] John Nolt. Free logic, *Stanford Encyclopedia of Philosophy*, Edward Zalta editor, 2018 `https://plato.stanford.edu/archives/fall2018/entries/logic-free/2018`.

[Nozick, 1968] Robert Nozick. Moral complications and moral structures. *Natural Law Forum*, 13:1–50, 1968.

[Oliver and Smiley, 2001] Alex Oliver and Timothy Smiley. Strategies for a logic of plurals. *Philosophical Quarterly*, 51:289–306, 2001.

[Parent and van der Torre, 2013] Xavier Parent and Leendert van der Torre. Input/output logics. In Dov Gabbay, John Horty, Xavier Parent, Ron van der Meyden, and Leon van der Torre, editors, *Handbook of Deontic Logic and Normative Systems*, volume 1, pages 499–544. College Publications, Milton Keynes, 2013.

[Parent, 2011] Xavier Parent. Moral particularism in the light of deontic logic. *Artificial Intelligence and Law*, 19:75–98, 2011.

[Parfit, 1986] Derek Parfit. *Reasons and Persons*. Oxford University Press, Oxford, 1986.

[Parfit, 1988] Derek Parfit. What we together do. Unpublished paper, 1988.

[Pinkert, 2015] Felix Pinkert. What if i cannot make a difference (and know it). *Ethics*, 125:971–998, 2015.

[Pollock, 1995] John Pollock. *Cognitive Carpentry*. MIT Press, Cambridge, 1995.

[Portmore, 2011] Douglas Portmore. *Commonsense Consequentialism: Wherein Morality Meets Rationality*. Oxford University Press, New York, 2011.

[Portmore, 2013] Douglas Portmore. Perform your best option. *Journal of Philosophy*, 110:436–459, 2013.

[Portmore, 2017] Douglas Portmore. Maximalism vs omnism about permissibility. *Pacific Philosophical Quarterly*, 98:427–452, 2017.

[Portmore, 2018] Douglas Portmore. Maximalism and moral harmony. *Philosophy and Phenomenological Research*, 96:318–341, 2018.

[Portmore, 2019] Douglas Portmore. *Opting for the Best*. Oxford University Press, New York, 2019.

[Prakken and Horty, 2011] Henry Prakken and John Horty. An appreciation of john pollockâĂŹs work on the computational study of argument. *Argument and Computation*, 3:1–19, 2011.

[Prakken, 2005] Henry Prakken. A study of accrual of arguments. In *Proceedings of Tenth International Conference on Artificial Intelligence and Law*, pages 85–94, 2005.

[Prakken, 2010] Henry Prakken. An abstract framework for argumentation with structured arguments. *Argument and Computation*, 1:93–124, 2010.

[Ramsey, 1931 1923] Frank Ramsey. Truth and probability. In R.B. Braithwaite, editor, *The Foundations of Mathematics and other Logical Essays*, pages 156–198. Hartcourt Brace Company, New York, 1931 [1923].

[Rayo, 2002] Agustín Rayo. Word and objects. *Noûs*, 36:436–464, 2002.

[Rayo, 2007] Agustín Rayo. Plurals. *Philosophy Compass*, 2:411–427, 2007.

[Raz, 2002 1975] Joseph Raz. *Practical Reasoning and Norms*. Oxford University Press, Oxford, 2002 [1975].

[Rechenauer and Roy, 2014] Martin Rechenauer and Olivier Roy. The logical structure of Scanlon's contractualism. In *Deontic Logic and Normative Systems: DEON 2014*, pages 166–176, 2014.

[Regan, 1980] Donald Regan. *Utilitarianism and Co-operation*. Oxford University Press, New York, 1980.

[Reiter, 1980] Raymond Reiter. A logic for default reasoning. *Artificial Intelligence*, 13:82–132, 1980.

[Roberts, 2015] M.A Roberts. The non-identity problem, 2015.

[Ross, 1930] W. D. Ross. *The Right and the Good*. Oxford University Press, Oxford, 1930.

[Ross, 2010] Jacob Ross. The irreducibility of personal obligations. *Journal of Philosophical Logic*, 39:307–327, 2010.

[Ross, 2012] Jacob Ross. Actualism, possibilism, and beyond. In Mark Timmons, editor, *Oxford Studies in Normative Ethics*, volume 2, pages 243–282. Oxford University Press, Oxford, 2012.

[Sayre-McCord, 1986] Geoffrey Sayre-McCord. Deontic logic and the priority of moral theory. *Noûs*, 20:179–197, 1986.

[Scanlon, 1998] Thomas Scanlon. *What We Owe to Each Other*. Belknap Press, Cambridge, 1998.

[Scanlon, 2014] Thomas Scanlon. *Being Realistic About Reasons*. Oxford University Press, Oxford, 2014.

[Schein, 1993] Barry Schein. *Plurals and Events*. MIT Press, Cambridge, 1993.

[Schein, 2006] Barry Schein. Plurals. In Ernest Lepore and Barry Smith, editors, *The Oxford Handbook of Philosophy of Language*, pages 716–767. Oxford University Press, Oxford, 2006.

[Schroeder, 2007] Mark Schroeder. *Slaves of the Passions*. Oxford University Press, Oxford, 2007.

[Schroeder, 2009] Mark Schroeder. Means-end coherence, stringency, and subjective reasons. *Philosophical Studies*, 143:223–248, 2009.

[Schroeder, 2011a] Mark Schroeder. Holism, weight, and undercutting. *Noûs*, 45:328–344, 2011.

[Schroeder, 2011b] Mark Schroeder. Ought, agents, actions. *Philosophical Review*, 120:1–41, 2011.

[Schroeder, 2015a] Mark Schroeder. Hypothetical imperatives: Scope and ju-

risdiction. In Robert Johnson and Mark Timmons, editors, *Reason, Value, and Respect*, pages 89–100. Oxford University Press, Oxford, 2015.

[Schroeder, 2015b] Mark Schroeder. *Being Realistic About Reasons*, by T. M. Scanlon. *Australasian Journal of Philosophy*, 93:195–198, 2015.

[Schroeder, 2018] Mark Schroeder. Getting perspective on objective reasons. *Ethics*, 128:289–319, 2018.

[Sergot, 2013] Marek Sergot. Normative positions. In Dov Gabbay, John Horty, Xavier Parent, Ron van der Meyden, and Leon van der Torre, editors, *Handbook of Deontic Logic and Normative Systems*, volume 1, pages 353–406. College Publications, Milton Keynes, 2013.

[Setiya, 2014] Kieran Setiya. What is a reason to act. *Philosophical Studies*, 167:221–235, 2014.

[Sher, 2019] Itai Sher. Comparative value and the weight of reasons. *Economics & Philosophy*, 35:103–158, 2019.

[Sidgwick, 1981 1907] Henry Sidgwick. *The Methods of Ethics*. Hackett Publishing, Indianapolis, 7 edition, 1981 [1907].

[Slote, 1984] Michael Slote. Satisficing consequentialism. *Proceedings of the Aristotelian Society*, 58:139–163, 1984.

[Snedegar, 2013] Justin Snedegar. Negative reason existentials. *Thought*, 2:108–116, 2013.

[Snedegar, 2014] Justin Snedegar. Contrastive reasons and promotion. *Ethics*, 125:39–63, 2014.

[Snedegar, 2017] Justin Snedegar. *Contastive Reasons*. Oxford University Press, Oxford, 2017.

[Sobel, 1976] John Howard Sobel. Utilitarianism and past and future mistakes. *Noûs*, 10:195–219, 1976.

[Tempkin, 2015] Larry Tempkin. *Rethinking the Good*. Oxford University Press, Oxford, 2015.

[Timmerman and Cohen, 2016] Travis Timmerman and Yishai Cohen. Moral obligations: Actualist, possibilist, or hybridist. *Australasian Journal of Philosophy*, 94:672–686, 2016.

[Timmerman, 2015] Travis Timmerman. Does scrupulous securitism stand-up to scrutiny. *Philosophical Studies*, 172:1509–1528, 2015.

[Tucker, 2018] Dustin Tucker. Variable priorities and exclusionary reasons in input/output logic. *Journal of Philosophical Logic*, 47:947–964, 2018.

[Uzquiano, 2004] Gabriel Uzquiano. The supreme court and the supreme court justices. *Noûs*, 38:135–153, 2004.

[Uzquiano, 2011] Gabriel Uzquiano. Plural quantification and modality. *Proceedings of the Aristotelian Society*, 111:219–250, 2011.

[Valentini, 2012] Laura Valentini. Ideal vs. non-ideal theory. *Philosophy Compass*, 7:654–663, 2012.

[Verheij, 1995] Bart Verheij. Accrual of arguments in defeasible argumentation. In *Proceedings of the Second Dutch/German Workshop on Nonmonotonic*

Reasoning, page 217–224, 1995.

[Vessel, 2003] John-Paul Vessel. Counterfactuals for consequentialists. *Philosophical Studies*, 112:103–125, 2003.

[Vessel, 2009] John-Paul Vessel. Defending a possibilist insight in consequentialist thought. *Philosophical Studies*, 142:183–195, 2009.

[Vessel, 2016] John-Paul Vessel. Against securitism, the new breed of actualism in consequentialism. *Utilitas*, 28:164–178, 2016.

[von Wright, 1951] G.H von Wright. Deontic logic. *Mind*, 60:1–15, 1951.

[Way, 2017] Jonathan Way. Reasons as premises of good reasoning. *Pacific Philosophical Quarterly*, 98:251–270, 2017.

[Wedgwood, 2006] Ralph Wedgwood. The meaning of 'Ought'. In Russ Schafer-Landau, editor, *Oxford Studies in Metaethics*, volume 1, pages 127–160. Oxford University Press, Oxford, 2006.

[White, 2017] Stephen White. Transmission failures. *Ethics*, 127:719–732, 2017.

[Williams, 1981 1979] Bernard Williams. Internal and external reasons. In *Moral Luck*, pages 101–113. Cambridge University Press, Cambridge, 1981 [1979].

[Williams, 1981 1980] Bernard Williams. *Ought* and moral obligation. In *Moral Luck*, pages 114–123. Cambridge University Press, Cambridge, 1981 [1980].

[Williams, 1988 1965] Bernard Williams. Ethical consistency. In Geoffrey Sayre-McCord, editor, *Essays on Moral Realism*, pages 41–58. Cornell University Press, Ithaca, 1988 [1965].

[Williamson, 2008] Timothy Williamson. *Philosophy of Philosophy*. Blackwell Publishing, Oxford, 2008.

[Williamson, 2010] Timothy Williamson. Necessitism, contingentism, and plural quantification. *Mind*, 119:657–748, 2010.

[Williamson, 2015] Timothy Williamson. *Modal Logic as Metaphysics*. Oxford University Press, Oxford, 2015.

[Woodard, 2008] Chris Woodard. *Reasons, Patterns, and Cooperation*. Routledge, London, 2008.

[Yi, 1999] Byeong-uk Yi. Is two a property. *Journal of Philosophy*, 96:163–190, 1999.

[Zimmerman, 1996] Michael Zimmerman. *The Concept of Moral Obligation*. Cambridge University Press, Cambridge, 1996.

[Zimmerman, 2006] Michael Zimmerman. The relevant risks to wrongdoing. In Kris McDaniel, Jason Raibley, Richard Feldman, and Michael Zimmerman, editors, *The Good, The Right, Life And Death: Essays in Honor of Fred Feldman*, page 151–172. Ashgate, Aldershot, 2006.

Shyam Nair
Arizona State University, USA
Email: gsnair@asu.edu

9

Logic and the Law: Philosophical Foundations, Deontics, and Defeasible Reasoning

Guido Governatori, Antonino Rotolo and Giovanni Sartor

Abstract. This chapter is a light-weighted overview of significant contributions to legal logic insofar as they involve deontic reasoning and related methods. A special emphasis is given to defeasible reasoning, which has been the major topic for legal reasoning in the last decades. The chapter is divided into three parts and the layout is as follows. Part 1 provides an introductory outline. In particular, we briefly recall an issue that was discussed in the context of deontic logic and that has been as well a hot research theme in legal reasoning, i.e., the very possibility of the use of logic in the law. Part 2 reconstructs the contribution of the literature about some classic topics or methods in deontic logic as relevant for the law: normative positions, the concept of permission, contrary-to-duty reasoning, input/output logic, algebras for normative systems, norm change, defeasibility in law. Part 3 is the largest one and offers, from our previous work, a unifying formal framework, based on Defeasible Logic, re-addressing some of the topics that we have already discussed in Part 2: legal hierarchies and dynamics, institutional agency and normative positions, and deontic aspects of legal interpretation.

1 Preface **659**

2 The possibility of logic in the law **661**

3 Introduction **664**

4 Basic concepts from legal theory **664**
4.1 Legal provision, legal norm, legal judgement 664

This work was partially supported by EU H2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No. 690974 for the project *MIREL: MIning and REasoning with Legal texts*

Handbook of Deontic Logic and Normative Systems, Volume 2

4.2 Norms . . . 665
4.3 Normative judgements . . . 668
4.4 Normative positions . . . 669
4.4.1 Deontic judgements . . . 669
4.4.2 Potestative judgements . . . 672

5 Normative systems 674
5.1 Logics and algebras for normative systems . . . 676
5.2 Concepts of permission in legal systems . . . 678
5.3 Contrary-to-duty reasoning and the law . . . 681
5.3.1 The $\otimes$ logic and its semantics . . . 683
5.4 Normative dynamics in the law . . . 686

6 Defeasibility in legal reasoning 688
6.1 Meanings of 'defeasibility' in the law . . . 688
6.2 Defeasibility and argumentation layers in the law . . . 690

7 Introduction 694

8 Basic defeasible logic 695
8.1 Further readings . . . 700

9 Defeasible deontic logic 701
9.1 Further readings . . . 704

10 Modelling permissions 705
10.1 Permissions and defeasibility . . . 705
10.2 Permissions, obligations, and preferences . . . 707
10.3 Further readings . . . 715

11 Institutionalised agency and normative positions 715
11.1 Introduction . . . 716
11.2 The framework and possible developments . . . 718
11.3 Further readings . . . 723

12 Legal provisions and legal norms: Interpretation and deontics 724
12.1 Introduction . . . 724
12.2 Deontic defeasible logics for legal interpretation . . . 728
12.3 Further readings . . . 733

13 Defeasible deontic logic with time **733**
13.1 Basics 733
13.2 From rules to meta-rules 736
13.3 Further readings 740

14 Modelling legal changes **741**
14.1 Types of legal change 741
14.2 Modifications of scope: Derogation 742
14.3 Textual modifications: Substitution 743
14.4 Temporal modifications 744
14.5 Modifications on norm validity and existence: Annulment vs. abrogation 745
14.6 Intermezzo – temporal dynamics and retroactivity 747
14.7 Modifications on norm validity and existence: Annulment vs. abrogation (cont'd) 748
14.8 Further readings 750

15 Conclusion **750**
15.1 Further readings 754

Part I
Introductory Outline

1 Preface

The relationship between logic and law has been a troublesome one and it has been object of much philosophical debate in the past century (cf. [Horovitz, 1972]). Several scholars have denied the usefulness of logical methods in law and legal theory, while others have strongly argued in favour of a logic-driven analysis of law and legal reasoning (e.g., [Kalinowski, 1959]). Be it as it may, this latter view has generated decades of interesting work at the interface of law, logic but also philosophy and artificial intelligence, and this work is the object of the present chapter.

H. L. A. Hart, among other legal philosophers, clarified the roles, activities, and functions that systems of norms and normative reasoning play in the law [Hart, 1994]:

Norm recognition and hierarchies: legal systems provide criteria for establishing whether norms belong to them; also, legal systems assign to their norms a different ranking status and organise them in hierarchies;

Norm application: legal norms are applied to concrete cases and legal systems include criteria to correctly apply their norms to such cases;

Norm change: legal norms and systems change and legal systems identify criteria governing their dynamics.

In a reasoning perspective, we can think of the above functions in terms of logical methods for modelling reasoning about

- norm types (e.g., when different reasoning methods are needed for handling different norms, such as regulative and constitutive norms), the structure of legal systems (e.g., when some norms are precondition for inferring or issuing other norms), and how norms are related with one another (e.g., when one norm overrides another one in case of conflicts) (**Norm recognition and hierarchies**);
- the interpretation of legal provisions (e.g., when different interpretive canons, as applied to the same provision, offer different legal solutions for a concrete case) and the application of legal norms (e.g., when norms are used to draw conclusions defeasibly, since possible exceptions can apply) (**Norm application**);
- how norms and legal systems are revised (**Norm change**).

Theoretically speaking, legal logic is thus a discipline that revolves around all the above three aspects of the law.

The present chapter is devoted to a short and light-weighted overview of some significant contributions on those issues. The overview should by no means be considered exhaustive and just briefly considers some key issues selected by the authors. In particular, this chapter is meant to offer an overview on topics falling within **Norm recognition and hierarchies** and **Norm change**, where deontic reasoning typically plays a role. In the context of norm application deontics has a limited role, except for some general aspects, such as interpretation, which will be briefly discussed in Section 12. However, we will devote large part of the chapter on defeasibility, which is of paramount importance in **Norm application**. A general outline on law and logic covering also norm application can be found in [Prakken and Sartor, 2015]. Notice that some classical approaches to deontic logic (such as algebras for normative systems, the concept of permission, and contrary-to-duty reasoning), if

used in the legal domain, contribute to eliciit the nature and structure of legal systems, thus being relevant for the issues **Norm recognition and hierarchies** and **Norm application**. The reader will find some sections in this chapter that address classical topics of deontic logic: our purpose is not to treat them comprehensively—other chapters in this handbook do so—but rather to handle those issues that are directly relevant for the law.

Chapter Layout The chapter is divided into three parts and the layout is a follows. Part 1, which includes this preface, provides an introductory outline. In particular, we briefly recall an issue that was discussed in the context of deontic logic and that has been as well a hot research issue in legal reasoning, i.e., the very possibility of the use of logic in the law. Part 2 reconstructs the contribution of the literature about some classic topics in legal reasoning which involve deontics but go beyond the Standard Deontic Logic paradigm: Section 4 introduces some fundamental concepts from legal theory such as norm, normative judgement, and normative position, Section 5 consider various issues related to the concept of normative system, such as permission, contrary-to-duty obligation, norm change, and so forth, Section 6 illustrates the meaning of defeasibility in law. Given the topics of Part 2, Part 3 offers, from previous work, a unifying formal framework, based on Defeasible Logic, addressing the following topics: basic logics for legal reasoning (Section 8), defeasible deontic reasoning (Section 9), defeasible permissions (Section 10), institutional agency and normative positions (Section 11), temporal normative reasoning (Section 13) and legal hierarchies and dynamics (Section 14), and and deontic aspects of legal interpretation (Section 12).

2 The possibility of logic in the law

The use of formal logic in the law has been the object of countless criticisms, which can be classified into two distinct categories. A first group of objections, which we could define *radical objections*, has to do with the presumed impossibility to apply the logic to normative reasoning in general. A second level of criticism, which we can instead define *moderate*, sees logic as a method only capturing a few (and easy) aspects of legal reasoning.

The *first category of criticism*, in effect, can undermine the root of any research which aims at using formal logic as an instrument of conceptual

clarification of legal analysis and reasoning. In this sense, as [Alchourrón and Bulygin, 1981] have recalled, two fundamental problems have to be addressed if logic should be applied to norms. On the one hand, it is claimed that no logic of norms can properly exist because norms do not have any truth value, in contrast with descriptive statements.[1] On the other hand, [Kelsen, 1979], among others, argued that the logic of legal norms is in general groundless because there is no significant logical relationship among legal norms: in fact, norms are valid, and their validity is a qualification which is *constitued* by the act of will of those subjects that create or apply them. Kelsen maintained that the validity of any legal norm n_0 can only be based on the validity of another norm that empowers the subject enacting n_0 to issue it; it cannot result from the derivability of n_0 from more general norms through logically correct inferences. In this sense, the production of n' is based on the will of a subject empowered by the legal system.

As regards *the first type of criticism*, many counterarguments have been presented in the literature, which have attempted to circumvent the obstacle. Some have argued that logic should not necessarily be applied only to sentences susceptible of truth or falsity, but that, otherwise, it mainly works on a syntactic notion of consequence relation [Alchourrón and Martino, 1990]. In addition, if semantics is needed, a dichotomy of the type $(1, 0)$ is sufficient (denoting any dichotomy such as valid/invalid, issued/unissued, just/unjust, etc) [Kalinowski, 1953]. Others have pointed out that there is no genuine logic of norms, but that it is however possible to develop a logic of normative propositions, or rather of descriptive propositions about norms [von Wright, 1963]. Some have instead defended the existence of an ideal normative dimension to which the norms could somehow match, and thus be qualified *latu sensu* as true or false [Kalinowski, 1972].

Other counterarguments can be mentioned. For example, consider the following. As is well-known, if we approach logical deduction semantically, we are usually able to focus on a very important property of it, i.e., the fact that deduction is truth-preserving: an argument is deductively valid if there is no interpretation (in the semantics) in which its premises are all true and its conclusion false. Clearly, such a semantic cannot apply to norms, if norms cannot have truth values. However, one may affirm that an logical inference with norms does not transmit the truth to the

[1] A classic formulation of such a skepticism is the so-called Jørgensen dilemma [Jørgensen, 1937 1938]. Some arguments supporting this thesis, which very popular among legal philosophers, have been offered by Kelsen, 1979 and by von Wright, 1963. For a more recent formal analysis of this topic, see [Makinson, 1999a].

conclusion of the premises (*truth-preservation*), but it rather preserves validity (*validity-preservation*) [Ross, 1968]. A stronger position is the one defended by Ota Weinberger, who admits that the logic of norms must have its own specificity with respect to the logic applied to descriptive utterances, but who also is highly critical towards those who, like Kelsen, believe that we cannot conceptualise logical relations between legal norms [Weinberger, 1981; Weinberger, 1989]. According to Weinberger, in fact, validity-preservation is not the only qualification of relations between the norms. A norm n is valid in the legal system S iff n is enacted by a subject that is authorised in S to enact a certain class of norms, then the logical principle of subsumption must apply at any rate, otherwise it is not possible to "apply" a power-conferring norm to empower this subject. Of course—Weinberger argues—this argument is not conclusive except in regard to (in Kelsen's view) the presupposed basic norm, which has not been enacted. However, overall, logic is needed precisely to make this basic norm applicable, thus providing the foundations for the entire legal system.

The second category of criticism—i.e., that logic only captures a few aspects of legal reasoning (see, for a broad discussion, among others [Alexy, 1989; Peczenik, 1989; MacCormick, 2005])—was addressed in the logic literature by broadening the scope and the techniques of legal logic. Indeed, for a long time the criticism was directly linked to the limits of judicial syllogism in reconstructing judicial decisions and reasoning. Of course, logics go beyond syllogistic inferences, thus offering new tools and methods—covering also the dialectical aspects of argumentation (cf. [Prakken and Sartor, 2015])—over a large variety of issues in the law.

In essence, the remainder of this chapter offers some answers to this second type of criticisms.

Part II
Deontic Reasoning in the Law – Classical Glimpses Beyond Standard Deontic Logic

3 Introduction

For several decades, most efforts were devoted to study deontic logic as a branch of modal logic. Among these efforts, themes like Standard Deontic Logic and its limits, or alternatives like the Andersonian-Kangerian reduction have been in the agenda of many deontic logicians (see Chapters 1, 3, 4, 5 and 7 of this handbook).

Recent developments of deontic logic that go beyond Standard Deontic Logic and its discussion are extensively reported on in other chapters of this handbook. If we move to the legal domain, we can reconsider some of them, as they were partly motivated by specific problems in the law.

This section briefly addresses relevant issues in deontic reasoning as applied to the legal domain.[2]

4 Basic concepts from legal theory

4.1 Legal provision, legal norm, legal judgement

According to contemporary legal theory and philosophy, there is usually no legal norm without interpretation. A norm practically is equivalent with one or more provisions plus the activity of their interpretation.[3] Hence:

Definition 4.1. *[Provision vs Norm] A* legal provision *p is an authoritative legal text within a given legal system. A* legal norm *n is the result of the interpretive process of one or more legal provisions $p_1, \ldots, p_n$ (see, e.g., [Peczenik, 1989]).*

[2]Some sub-sections elaborate on parts of [Grossi and Rotolo, 2011].

[3]Legal and social theorists are of course aware of exceptions, i.e., legal norms without corresponding provisions: canonical examples are from customary law, where norms are not positively stated—namely, textually formulated—by any formal authorities.

Once it is clear that legal provisions are different from norms, a basic classification of legal concepts, which possibly result from legal texts, includes two main classes:

- Norms,
- Normative judgements, which concern typically (but not only) the effects of the application of legal norms.

In the remainder we recall the ways in which these concepts can be analysed and how they can be further classified.

4.2 Norms

Norms are propositions stating normative judgements. Norms can be unconditioned, that is their judgement may not depend upon any antecedent condition (consider, for example, the norm "everyone has the right to express his or her opinion"). Usually, however, norms are conditioned.[4]

The framework proposed by [Sartor, 2005] distinguishes between two types of conditioned norms: *rules*, which make a normative judgement dependent upon defeasibly sufficient conditions[5] and *factor links*, which make a normative judgement dependent upon contributory conditions.[6] However, the literature proposing norm classifications is very rich (among many others, [von Wright, 1963; Nino, 2013; Sartor, 2005]).

A fundamental distinction is made by some theorists, such as Ronald

[4]We should also consider the antecedents of conditioned norms, and introduce the traditional classification between juridical fact, acts (facts relevantly determined by humans), and declarations of will or intentions [Sartor, 2005; Pattaro, 2005]. In this way we might also characterise the notion of a source of law, by which we mean any fact that embeds normative propositions and makes them legally valid by virtue of such an embedment. Some sources of the law are events (like the issuing of a high court decision), while others are state of affairs (like the practice of a custom or a result declaration). In this way, even an unconditioned norm (closely similar to a normative judgement) can be conditioned the the state of affairs or event that produced it.

[5]Of course, the conditions listed in a legal provision are sometimes necessary for a given normative judgement (usually expressed in a converse conditional).

[6]While the notion of rule is largely known in many scientific communities and is close to intuition, the idea of factor link is somehow technical and developed within the AI&Law community. The intuition behind this idea is that it is not always possible to formulate precise rules, even defeasible ones, for aggregating the factors relevant for resolving a legal issue. For example: "The educational value of a work needs to be taken into consideration when evaluating whether the work is covered by the copyright doctrine of fair use". Of course, when a factor contributes to undercut other rules, namely, to make them inapplicable, it can be called an exclusionary reason in a broad sense [Raz, 1990], even though Raz's orginal definition consider exclusionary reasons as "a second-order reason to refrain from acting for some reason" [Raz, 1990, p. 39].

Dworkin and Robert Alexy, between *rules* and *principles* (for an overview, see [Hage, 1997; Sartor, 2005]). While rules apply to cases according to an all-or-nothing logic, principles express legal values or fundamental rights; they

> "are norms which require that something be realized to the greatest extent possible given the legal and factual possibilities. Principles are optimization requirements, characterized by the fact that they can be satisfied to varying degrees, and that the appropriate degree of satisfaction depends not only on what is factually possible but also on what is legally possible." [Alexy, 2002, p. 47]

Rules, instead,

> "are norms which are always either fulfilled or not. If a rule validly applies, then the requirement is to do exactly what it says, neither more nor less." [Alexy, 2002, p. 48]

While it has been disputed that this distinction makes sense from the logical point of view[7], it is still an interesting research issue to explore the formal role of values and principles in legal reasoning [Sartor, 2010].

Other classifications are proposed in literature, most of them motivated by the fact that different normative judgements follow from the application of norms. However, we can also identify different mechanisms and functions for norms, which are relatively independent of the legal effects resulting from them. In particular three types of norms have been identified (see [Rubino *et al.*, 2006]),

Definition 4.2 (Norm types)**.** *We can distinguish the following types of legal norms*

initiation norms, *that is, norms stating that a certain normative proposition starts to hold when the rule's conditions are satisfied. An example is "if one causes a damage, one has to compensate it";*

termination norms, *that is, norms stating that a normative proposition ceases to hold when the rule's conditions are satisfied. An example is "if one pays a debt, the obligation terminates";*

supervenience norms, *that is, norms stating that a normative proposition holds as long as the conditions the conditions are satisfied. An example is "if one is in a public office, one is forbidden to smoke".*

[7] For instance, [Sartor, 1995] argued that principles are nothing but rules with high degree of defeasibility.

Other important basic taxonomies include the distinction between *regulative* and *constitutive* norms, which was initially proposed in philosophy (cf., among others, [Rawls, 1955; Searle, 1969]):

> "[R]egulative rules regulate antecedently or independently existing forms of behaviour [...]. But constitutive rules do not merely regulate, they create or define new forms of behaviour. The rules of football or chess, for example [...] create the very possibility of playing such games." [Searle, 1969, p. 33]

> "A marriage ceremony, a baseball game, a trial, and a legislative action involve a variety of physical movements, states, and raw feels, but a specification of one of these events only in such terms is not so far a specification of it as a marriage ceremony, baseball game, a trial, or a legislative action. The physical events and raw feels only count as parts of such events given certain other conditions and against a background of certain kinds of institutions. Such facts as are recorded in my above group of statements I propose to call institutional facts. They are indeed facts; but their existence, unlike the existence of brute facts, presupposes the existence of certain human institutions. [...] These 'institutions' are systems of constitutive rules. Every institutional fact is underlain by a (system of) rule(s) of the form 'X counts as Y in context C.' Our hypothesis that speaking a language is performing acts according to constitutive rules involves us in the hypothesis that the fact that a man performed a certain speech act, e.g. made a promise, is an institutional fact." [Searle, 1969, p. 51–52]

The idea of constitutive rule has been subsequently imported in legal theory (see, e.g., [Ruiter, 2001]) and in legal logic (see, e.g., [Jones and Sergot, 1996; Grossi and Jones, 2013a]) to model, e.g., the concept of institutionalised power and the one of power-conferring norm.

Notice that different logical characterisations of constitutive rules are possible, stating, for example, that the conditions in the antecedent must occur in order to initiate the occurrence of the consequent. In particular, [Jones and Sergot, 1996] developed an analysis of the notion of institutionalised power by introducing a new conditional connective '$\Rightarrow_s$'. This connective expresses the counts-as connection holding in the context of an institution s. In short, this approach is roughly in line with Goldman's theory of actions generating actions [Goldman, 1970]. In this

perspective, it was argued that the generation of institutional facts via constitutive rules is quite close to the idea of a causal relation—contrary to [Searle, 1995]'s argument according to which the counts-as link rather amounts to a classificatory relation—and consequently some well-known axiom schemata, such as $A \Rightarrow_s A$, do not hold. Another formalisation, though openly inspired by Jones and Sergot, proposed some substantial changes in the light of a different philosophical interpretation of the counts-as relation [Gelati *et al.*, 2004]. Counts-as rules are meant to capture the constitutive character of institutional ontology and express institutional taxonomies. Accordingly, their function is to represent the constitutive ingredients of institutional facts, whose nature is conceptually distinct from that of the empirical facts.

4.3 Normative judgements

Let us consider the concept of legal judgement[8].

Definition 4.3 (Legal judgements)**.** *A* legal judgement *is the propositional constituent expressing a legal fact*[9]*. Legal judgments can be classified into the following kinds [Rubino* et al.*, 2006]:*

evaluative, *which indicates that something is good or bad, is a value to be optimised or an evil to be minimised (for example, "human dignity is a value", "participation ought to be promoted");*

qualificatory, *which ascribes a legal quality to a person or an object (for example, "x is a citizen", "x is an intellectual work", "x is a technical invention");*

definitional, *which specifies the meaning of a term (for example "x means y" or "by x it is meant y");*

deontic, *which imposes the obligation or confers the permission to do a certain action (for example "x has the obligation" or "x has the permission to do A");*

potestative, *which attributes powers (for example "a worker has the power to terminate his work contract");*

[8]Parts of this section recall and elaborate texts from [Rubino *et al.*, 2006].

[9]The general notion of legal fact is controversial (cf., among others [Greenberg, 2004]). Here, we can assume that a legal fact is a fact normatively qualified, i.e., a fact to which legal effects are attached, or giving rise to the occurrence, change or termination of any legal relationship, legal capacity, legal competence, subjective rights, or legal obligation.

evidentiary, *which establishes the conclusion to be drawn from certain evidence (for example "it is presumed that dismissal was discriminatory");*

existential, *which indicates the beginning or the termination of the existence of a legal entity (for example "the company ceases to exist");*

norm-concerning judgements, *which state the modifications of provisions or norms such as abrogation, repeal, substitution, and so on, or how other norms should be applied or interpreted.*

In the following we briefly report on the so-called theory of normative positions, which traditionally deals with the concepts of deontic judgement and potestative judgement.

4.4 Normative positions

4.4.1 Deontic judgements

Basic deontic modalities are those that usually are modelled in deontic logic in the form of expressions like $\mathsf{O}a$ (a is obligatory). In fact, a deontic judgement expresses the fact that a certain content is qualified by deontic modalities, typically obligation, prohibition and permission. Deontic concepts can be reduced to those of obligation and permission. Classically, prohibition can be defined in terms of the notion of obligation ($\mathsf{O}\neg$). *Basic deontic modalities* correspond to the standard deontic qualifications in deontic discourse. They are also called undirected deontic modalities, as no explicit reference is made to any subject which may be the beneficiary of the deontic qualification.

A well-known and relevant distinction for the legal domain is the one classifying deontic modalities into *ought-to-be* and *ought-to-do* judgements: the former express deontic qualifications whose content are states of affairs without necessarily mentioning actors or actions bearing relations with such states of affairs; the latter may be interpreted as expressing deontic qualifications of explicit actions. Although in many cases ought-to-be statements can be reframed as ought-to-do statements, it is quite controversial that this can be done in general. In fact, ought-to-be statements are often made when it is not known who will have the responsibility of realising the state of affairs though it is known that somebody has this responsibility (for a survey in the legal domain, see [Sartor, 2005]). An example of normative judgement involving an undirected ought-to-be qualification is the following: "The balance of a

bank account ought to be non-negative". An example of normative judgement involving an undirected ought-to-do qualification is the following: "Everybody has the obligation to pay taxes".

Another important distinction can be drawn for the law between *weak and strong permissions*: "An act will be said to be permitted in the weak sense if it is not forbidden; and it will be said to be permitted in the strong sense if it is not forbidden but subject to norm", i.e., when a norm explicitly states it is permitted [von Wright, 1963, p. 86]. This distinction may be crucial in characterising notions such as those of authorisation and derogation [Boella and van der Torre, 2005; Sartor, 2005; Stolpe, 2010b]. See Section 5.2 for a discussion.

Normative judgements stating *directed deontic modalities* can indicate the bearers (as originally proposed by seminal works such as [Herrestad and Krogh, 1995]) or the beneficiaries of the deontic qualifications specified in such judgements [Herrestad and Krogh, 1995; Gelati *et al.*, 2004; Sartor, 2005]. Contributions in this field also come from the so-called Kanger-Lindahl theory of normative positions, which has developed into a rich blend of modal deontic and action logics able to formalise a complex array of deontic and legal concepts.[10] Here, we focus on classifications taking into account beneficiaries only (for a comprehensive overview, cf. [Sartor, 2005]).

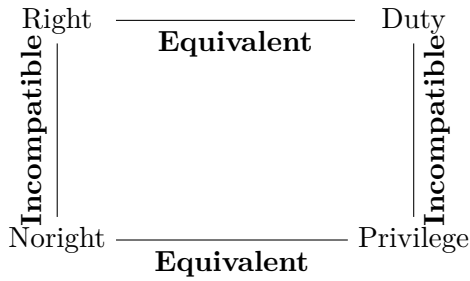
We distinguish two ways in which the indication of beneficiaries can take place: either the deontic qualification holds towards specified individuals, in which case we speak of an individualised qualification, or it holds towards everybody, in which case we speak of an *erga-omnes* qualification. An example of normative judgement involving a directed *erga-omnes* ought-to-be qualification is the following: "Traffic ought to be reduced in the interest of the every Italian citizen". An example of normative judgement involving a *directed individualised* ought-to-do qualification is the following: "In the interest of Mr. Jones, Ms. Smith has the obligation to pay him one thousand euros". An example of normative judgement involving a directed *erga-omnes* ought-to-do qualification is the following: "In the interest of the owner everybody is forbidden to use his/her property without his/her consent".

Directed obligative ought-to-do are also called *obligative rights*. Agent k has the obligative right that j does A iff it is obligatory, towards k, that j does A. An example of obligative right is "it is obligatory, towards Mary, that Tom pays 1,000 euros to John". Another type of obligative

[10] See [Kanger and Kanger, 1966] for an early exposition, [Sergot, 2001] for a reference treatment, and Chapter 5, Volume I, of this Handbook for a recent and comprehensive analysis of the theory.

rights are the exclusionary rights which concern the prohibition against performing certain inferences (against reasoning in certain ways), or against using certain kinds of premises for certain purposes, in the interest of a particular person. This is especially the case with anti-discrimination rules. For instance, in many legal systems employers are prohibited from adopting any decision having a negative impact on their employees on the basis of race or sex, and this prohibition, though also serving some collective purposes, is primarily aimed at promoting the interest of the employees in question.

Let us now specifically consider how we can conceptualise the difference between directed ought-to-do deontic judgements having a positive or a negative content, that is, concerning actions or omissions. Both obligations and permissions can be divided into positive and negative according to whether they concern an action or an omission. *Directed negative permissions* constitute what is also called *privilege* in the Hohfeldian language [Hohfeld, 1913; Hohfeld, 1917]: j has a privilege towards k, with regard to action A, iff it is permitted towards k that j omits to do A. Following again Hohfeld, we may use the less controversial expression *noright* to express that one does not have the obligational right that another does a certain action, that is, to denote the situation when the latter is permitted towards the former to omit that action. Therefore, we can say that k has a noright that j does A iff j is permitted, towards k, to omit A. Let us make an example both for privileges and norights. Assume for instance that Mary, a writer, has made a contract with Tom, a publisher, and has committed herself to write a novel for him. Mary's privilege would consist in Mary having permission towards Tom not to write the novel, a normative situation which could also be described as Tom's noright that Mary writes the novel. Hohfeld's analysis of these concepts is illustrated by what is sometimes called the first Hohfeldian square:



Positive and negative permissions can be merged into the concept of

faculty (for instance, by saying that a woman has the faculty of wearing a miniskirt when going to work, we mean that it is permitted both to wear it and not to wear it). When, for the benefit of a person, this person is both permitted to perform and to omit an action—that is, when the action is facultative—we can say that he or she has a *liberty right* with regard to that action. This notion can be further developed according to the fact that others (or the government) may have, always in the interest of that person, a prohibition to prevent the facultative action, and they may even have the obligation to provide means for its performance. This leads us to distinguish three kinds of liberty rights: a *mere liberty right*, a *negatively protected liberty right*, and a *positively protected liberty right*. In general we speak of a right to characterise the situation where a normative judgement is intended to benefit a particular person.

According to this notion of a right, the directed obligations of agent j for the benefit of agent k can be viewed as k's right, namely as k's obligative right towards j. The negation of a directed obligation is a directed permission. However, it counts as a right, namely, a permissive right, only when such negation is aimed at benefitting the author of the permitted action.

Another notion of a right is that of *liability rights*[11]. That j has a liability right concerning k's action A means that if k performs the permitted action A then k will have to perform another action B for the benefit of j. For example, consider a copyright regime when one is permitted to reproduce a protected work, but the author is entitled to a royalty for the reproduction of his or her work. In this case we have a normative connection between a permitted action and an obligation of the agent, to the benefit of another. However, for us this kind of legal position represents a conditional, namely, a norm, rather than a normative judgement.

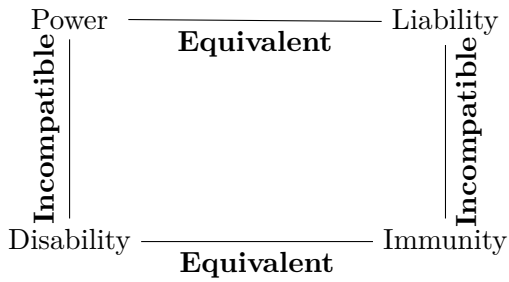
4.4.2 Potestative judgements

Potestative judgements concern the attribution of powers. The first level of classification proposed by [Sartor, 2005; Rubino *et al.*, 2006; Gelati *et al.*, 2004] includes the categories *Hohfeldian powers*, *enabling powers* and *declarative powers*: the first covers any action which determines a legal effect, the second only cases when the law aims at enabling the agent to produce the effects in this way, the third the case when the

[11] Despite the common term, notice that Hohfeld's concept of liability differs from the idea of liability right.

effect is produced through the agent's declaration of it. In more detail, we say that j has the declarative power to realise A to mean that if j declares A, then it is legally valid that A. For example, if x has the declarative power to terminate y's obligation towards x to do then if x declares that y's obligation towards x finishes, then it is legally valid that this obligation finishes.

The second Hohfeldian set includes *immunities*, *action powers*, *subjections* (the normative position that Hohfeld denotes as liability), and *disabilities*. Agent k has an *immunity* towards j with regard to the creation of position *Pos* for k, exactly if is not the case that j has that power. An *action-power* consists in a generic power to produce a legal effect through an action determining it. That k is in a state of *subjection* towards j, with regard to normative position *Pos*, means that j has the abstract enabling-power of determining *Pos* for k. For instance, debtor k is subject to creditor j in relation to j's power of freeing k from j's obligation. Agent j has a *disability* towards k, with regard to the creation of position *Pos* exactly if it is not the case that j has the abstract enabling power of creating *Pos* for x. Here below is a pictorial representation of the second Hohfeldian square:



A special kind of enabling powers, called *potestative rights*, can be distinguished, that is, powers which are meant to benefit the holder of the power. For example, if an animal y does not belong to anybody, then x has the potestative-right to start x's ownership of y, by capturing y. Notice that this power means that finding and capturing an animal puts x in the position of an owner toward everyone else, which is a complex notion—a multital right—merging many bilateral rights (both claim-rights and powers) toward everybody else.

While the formal analysis of the first Hohfeldian square could build on deontic logic as the underlying framework for a logic of obligation, a formal analysis of the second square appears more complex and challenging. The quest for such analysis was programmatically set by [Jones and

Sergot, 1996], a paper that sparked an interesting line of research at the interface of logic, philosophy and artificial intelligence in the last twenty years.[12]

The issue addressed by [Jones and Sergot, 1996] consists in the formal characterisation of a notion of *legal power* as involved in sentences such as "the president has the power to declare a state of emergency". This notion of power is viewed as grounded in the so-called constitutive rules, viz. legal rules such as "18 years of age counts as age of majority" or "the president's signature counts as the enactment of the bill". For instance, the latter rule establishes that the president has the power to enact a legislative bill. As extensively argued for instance in [Searle, 1995], these rules—often called counts-as conditionals—represent the basic brick of complex institutions such as legal systems, and [Jones and Sergot, 1996] developed a first logical analysis of them.

Several aspects of this topic has been treated in the literature (see Chapters 5 and 6, Volume I of this Handbook), but a logical and comprehensive account of legal normative positions is still missing (for an extensive discussion of open problems and a semi-formal analysis, see [Sartor, 2005]).

5 Normative systems

Most of the work in deontic logic has focused on the study of the concepts of obligation, permission, prohibition and related notions, but little attention has been dedicated on how these prescriptions are generated within a normative system.[13] The general idea of norms is that they describe conditions under which some behaviours are deemed as 'legal'. In the simplest case, a behaviour can be described by an obligation (or a prohibition, or a permission), but often norms additionally specify what are the consequences of not complying with them, and what sanctions follow from violations and whether such sanctions compensate for the violations.

In this general perspective, a very influential contribution, which is complementary to the (modal) logic-based approaches to deontic logic, is the one sparked by [Alchourrón and Bulygin, 1971]. The key feature of this approach is to study norms—viewed as dyadic constructs connecting a fact to a deontic consequence—not as formulae in some logical language,

[12] See [Grossi and Jones, 2013b] for a comprehensive overview of this field of research.

[13] A normative system can be understood as a, possibly hierarchically structured, set of norms and mechanisms that systematically interplay for deriving deontic prescriptions in force in a given situation.

but rather as primitive ordered pairs $\langle condition, consequence \rangle$. A large number of such pairs would constitute an interconnected system called a *normative system.*

This general approach, hence, *clearly distinguishes norms from permissions/obligations* (a distinction that falls within the general analysis reported in Section 4): permissions and obligations are the effects (i.e., the conclusions) of the application of prescriptive norms.

Viewed as parts of a bigger system, norms are therefore holistically considered to be uninterpretable if taken in isolation—unlike in logical semantics—and they acquire meaning only by relating to other norms in the system. Thus, the principle behind this approach is "no logic of norms without attention to the normative systems in which they occur" [Makinson, 1999b]. This idea draws inspiration also from the pioneering works by [Stenius, 1963], and focuses on the fact that normative conclusions derive from norms as interplaying together in normative systems. Indeed, it is essential in this perspective to distinguish norms (such as the prescriptive and permissive ones) from normative positions, e.g., obligations and permissions [Boella *et al.*, 2009]: the latter ones are merely the effects of the application of norms. The focus falls then on the problem of normative reasoning and its most characteristic features, such as defeasibility (which is discussed later).

The basic idea behind this approach to modelling normative systems goes hand in hand with the thesis according to which norms do not bear truth-values, and hence that deontic logics do not actually deal with norms, but rather with normative propositions, i.e., statements to the effect that certain conclusions follow from existing norms. For instance, in this view, $\mathsf{O}A$ would actually mean something like "there exist norms commanding A".[14]

In what follows, we sketch, very briefly, the basic ideas behind two of logical methods that in recent years have taken up and developed the normative systems approach to the analysis of norms: input/output logics and normative systems algebras. Also we will briefly discuss three classical topics in deontic logic that are relevant for the law and that are central for modelling the relation between obligations, norms, and normative systems: specific logics and algebras for normative systems,

[14] As we have mentioned above, the problem of whether norms bear or not truth values is an old one in philosophy, and was put forth in modern times by [Jørgensen, 1937 1938]. The significance of the problem has recently been reemphasised in [Hansen *et al.*, 2007], and a new approach to the problem emerged from the view of norms as 'dynamic' operators—speech acts—modifying ideality orders. We will briefly come back to this latter point in Section 5.4.

the concept of permission, and the notion of contrary-to-duty obligations.

5.1 Logics and algebras for normative systems

One important attempt to holistically reason about normative systems is based on Input/output logics (IOL). IOLs are formalisms introduced in [Makinson and van der Torre, 2000] that have been applied to the study of normative systems in a long series of papers (e.g.,[Boella and Van der Torre, 2004]). Such systems are viewed as rule-based processes of manipulation of inputs (factual premises) into outputs (normative conclusions).

An extensive treatment of IOL is offered in Chapter 8 of this handbook. The key idea behind the application of IOL to normative systems consists in representing conditional norms simply as ordered pairs (a, b) where a represents the antecedent of the rule, and b its consequent, i.e., as conditionals: "if a then b" where a has factual content and b normative content, viz. an obligation or a permission. Typically, both a and b are taken to be formulae from propositional logic. Each set of such ordered pairs can be seen as an inferential mechanism which, given an input, determines an output based on those connections.

Various definitions can be given of how to produce the output on the basis of a set of pairs, and all consist in ways of closing the given set of pairs by adding new pairs in accordance to some principles, of which we give two very simple examples:

$$SI : \frac{(a, b)}{(a \wedge c, b)} \qquad CT : \frac{(a, b), (a \wedge b, c)}{(a, c)} \tag{1}$$

where SI stands for strengthening of the input—essentially an antecedent strengthening property—and CT stands for cumulative transitivity. Formally, given a set $NORM$ of pairs, a closure operation C defined in terms of some of the above principles, and a set of facts A, the output of $NORM$ given C and a set of input formulae I is:

$$out_C(NORM, A) \quad = \quad \{b \mid (a, b) \in C(NORM) \text{ and } b \in A\} \tag{2}$$

Intuitively, $NORM$ represent the norms of a normative system and C the principles according to which the system makes the norms interact with one another. As the reader might have already noticed, this represents a very high-level abstraction of the workings of a normative system. Depending on the (many) ways the output operation is defined, IOL can be used to capture very different principles for reasoning with norms (among which defeasibility). This modelling freedom brought IOL to be

applied not so much to the study and analysis of normative reasoning in *actual* normative systems, but rather to the specification of *artificial* normative systems in the field of artificial intelligence, see [Boella and Van der Torre, 2004].

Lindhahl and Odelstad [2000] advocate instead an algebraic analysis of normative systems. The approach is very close in spirit to the one, discussed above, of IOL. However, the formal machinery deployed is not based on logic, but rather hinges on several algebraic and order-theoretical notions. In this section, we provide just a brief sketch of the basic technical ideas underpinning the framework.

According to this approach, norms can be seen—exactly as in IOL—as simple pairs $\langle a, b\rangle$ connecting (factual) conditions to (normative) consequences. Both conditions a and consequences b are taken to be elements of a set X upon which a Boolean algebra $\langle X, \sqcap, -, \bot\rangle$ is defined. Within such a structure, the normative relation between condition a and consequence b is given by extending the preorder yielded by the algebra.[15] The idea is that while the preorder—let us call it $\preceq$—represents some form of logical implication, normative systems add on the top of it the possibility of drawing more conclusions by some form of 'legal' implication—let us call it ρ. In other words, each normative system introduces, by stipulation, a consequence relation which is stronger than the logical one: $\preceq \subseteq \rho$. The intuition is that, for instance, the fact that being obliged to pay taxes follows from having a paid job is not a matter of logic, but a matter of stipulation[16].

Therefore, in Lindahl and Odelstad's view normative systems can be studied as Boolean algebras supplemented by a binary relation ρ. This is, in a nutshell, the key idea behind the approach. Space limitation prevents us to provide more details. It should be mentioned, however, that [Lindahl and Odelstad, 2000] was followed by a number of papers developing an extensive theory of normative systems on the ground of the simple intuition we have sketched above.[17]

[15] A preorder can always be associated to a given Boolean algebra in the following way:

$$a \preceq b \quad \textit{iff} \quad a \sqcap b = a. \tag{3}$$

[16] As is well-known, the idea that legal effects do not follow from norms by logic but, rather, by stipulation was notably defended in legal theory by [Kelsen, 1991].

[17] An interesting contribution is, for instance, offered by [Lindahl and Odelstad, 2008]. A comprehensive overview is offered in Chapter 9, Volume I of this handbook.

5.2 Concepts of permission in legal systems

The concept of permission plays an important role in the law in that it is crucial in characterising notions such as those of authorisation and derogation [Boella and van der Torre, 2005; Sartor, 2005; Stolpe, 2010b]. For example, consider when we subscribe to an on-line sale agreement accepting to enter our personal data on the condition that this information is only used for shipping, and other necessary purposes to communicate with us or deliver the products to us. Here, the permission to use our personal data is an exception to a general prohibition.

Despite this fact, the concept of permission is still elusive in deontic logic and logicians for a long time have mostly overlooked it. The history of deontic logic offers however some well-known key ideas to interpret it. Indeed, the original intuition (proposed by [von Wright, 1951], among others) that permissions are the modal dual of obligations is technically simple and attractive. If permission is the dual of obligation, this means that $\mathsf{P}A \equiv \neg\mathsf{O}\neg A$, i.e., that something is permitted *iff* it is not prohibited. This view proved to be partial and simplistic (for a discussion, see [von Wright, 1963; Alchourrón and Bulygin, 1984; Alchourrón, 1993]), as it hardly allows us to grasp cases where a legal system explicitly states that some action is permitted: indeed, in this case that A is *explicitly* permitted *implies* that A is not prohibited, but not the other way around. (We present below two examples and some further intuitions.)

This is one of the reasons why the attempt to reduce permissions to duals of obligations has been criticised.

Hence, subsequent contributions enriched the picture in several directions. The distinction between *weak* (or *negative*) and *strong* (or *positive*) [von Wright, 1963] plays an important role in this regard. The former corresponds to saying that some A is permitted iff $\neg A$ is not provable (or it is false) as mandatory. In other words, something is allowed by a code only when it is not prohibited by that code. At least when dealing with unconditional obligations, the notion of weak permission is trivially equivalent to the dual of obligation [Makinson and van der Torre, 2003].

The concept of strong permission is more complicated, as it amounts to saying that some A is permitted by a code iff such a code explicitly states that A is permitted. It follows that a strong permission is not derived from the absence of a prohibition, but is explicitly formulated in a permissive norm. The complexities of this concept depend on the fact that, besides "the items that a code explicitly pronounces to be permitted, there are others that in some sense follow from the explicit

ones". The problem is hence "to clarify the inference from one to the other" [Makinson and van der Torre, 2003, p. 391–2]. For example, if some B logically follows from A, which is strongly permitted, is B strongly permitted as well?

Here below are examples illustrating the two concepts.

Example 5.1 (Weak permission – Taxation)**.** *Consider a legal system including one or more provisions, which are the only ones in the system governing double taxation, and whose joint interpretation makes true the following norm:*

> *If one lives in Italy for more than 183 consecutive days over a 12-month period, then she is obliged to pay taxes in Italy on her worldwide income.*

According to this system, if you lived in Italy for 60 consecutive days then it is weakly permitted for you not to pay your taxes in Italy.

Example 5.2 (Strong permission – U.S. Copyright Act)**.** *Section 504(c)(1) (Remedies for infringement: Damages and profits) of the U.S. Copyright Act (17 USC §504) states the following*[18]*:*

> *Except as provided by clause (2) of this subsection, the copyright owner may elect, at any time before final judgment is rendered, to recover, instead of actual damages and profits, an award of statutory damages [. . .]*

If we get closer to the literature on legal reasoning, [Alchourrón and Bulygin, 1981; Alchourrón and Bulygin, 1984] argued however that there is only one prescriptive sense of permission, while the distinction between weak and strong permission makes sense only at a descriptive level, depending on how any permission is obtained within a system of norms: typically—and somehow abusing terminology—the same permissive statement $\mathsf{P}a$ is either a strong or a weak permission depending on whether it has been obtained either through an explicit permissive norm or by directly denying that the opposite prohibition holds (for instance, by stating that relevant forbidding norms do not apply or are somehow blocked). Legal theorists such as Alf Ross and Norberto Bobbio [Ross, 1968; Bobbio, 1958] maintained that legal permissions are in fact exceptions to obligations imposing the opposite, even though this did not lead them

[18]This example shows a permissive norm expressing a peculiar type of strong permission, i.e., a permissive right, which is directed permissions aimed at satisfying an interest of the person being permitted [Sartor, 2005].

(Ross, in particular) to clearly link the concept of exception with the one of strong permission. Other theorists even denied the usefulness of seeing strong permissions as exceptions [Opalek and Wolenski, 1991; Royakkers and Dignum, 1997], since the former ones just express, in a different way, standard deontic indifference in normative systems. This objection by [Opalek and Wolenski, 1991; Royakkers and Dignum, 1997] was instead rejected by [Alchourrón and Bulygin, 1981].

Features such as the distinction between strong and weak permission show the multi-faceted nature of permission and permissive norms: a new interest has emerged in recent years on this topic [Makinson and van der Torre, 2003; Boella and van der Torre, 2003a; Boella and van der Torre, 2003b; Brown, 2000; Stolpe, 2010b; Stolpe, 2010a; Governatori *et al.*, 2013a]. Those contributions followed some, if not all the following principles:

1. the concept of permission concerns how permissive norms and other types of norms interact within systems;
2. a fundamental role of positive permissions is that of stating exceptions to obligations; hence, positive permissions are supposed to override or at least block some deontic conclusions coming from other norms;[19]
3. another fundamental role of positive permissions is to prevent a legislator from issuing future obligations;
4. the logical space of weak permission is the one left unregulated by mandatory norms.

[Makinson and van der Torre, 2003; Boella and van der Torre, 2003a; Stolpe, 2010b] all worked on Input/Output Logic [Makinson and van der Torre, 2000]. This logic allows for defining different concepts of permission [Makinson and van der Torre, 2003; Boella and van der Torre, 2003a].[20] First of all, ordered pairs are partitioned into obligation norms (G) and permissive norms (P). Thus:

Negative permission: (a, x) is a negative permission w.r.t. G iff $(a, \neg x) \notin out_C(G)$; if x is not prohibited by the system given a, then is negatively permitted under those factual conditions a.

Static permission: (a, x) is a static permission w.r.t. (G, P) iff $(a, x) \in out_C(G \cup \{(c, d)\})$ for some $(c, d) \in P$; (a, x) is statically permitted

[19] Under this reading, each positive permission works as a *Lex specialis* which derogates to a general and opposite obligation.

[20] For the notation, see Section 5.1 above. [Stolpe, 2010b] offers a different technical treatment, which is however in line with most intuitions discussed by [Makinson and van der Torre, 2003; Boella and van der Torre, 2003a].

iff it follows from adding a permissive norm to G;

Dynamic permission: (a, x) is a dynamic permission w.r.t. (G, P) iff $(c, \neg d) \in out_C(G, \cup\{a, \neg x\})$ for some $(c, d) \in P$; (a, x) is permitted when, given the obligations in G, we cannot prohibit x under the condition a without prohibiting d under condition c which is however explicitly permitted by the system.

Another concept of permission was proposed in [Stolpe, 2010b] to specifically capture the idea of exception[21]:

Exemption: (a, x) is an exemption w.r.t. (G, P) iff $(a, \neg x) \in out_C(G) \backslash out_C(G) - (c, \neg d)$ for some $(c, d) \in P$; (a, x) is an exception if the code contains the prohibition of x under condition a which, unless it is removed, it clashes with an explicit permission in P.

A different general contribution on the topic [Governatori *et al.*, 2013a] is based on Modal Defeasible Logic: we will describe it with some details in Section 10.

5.3 Contrary-to-duty reasoning and the law

One of the main research themes in deontic logic concerns reasoning with contrary-to-duty (CTD) obligations. These are obligations that are triggered by the violation of other obligations. E.g., "you ought not to kill, but if you kill it is obligatory that you are punished". Roughly, contrary-to-duty obligations have to do with sub-ideal obligations. Clearly, this is crucial for the law, which does not only provide prescriptive statements, but among other things, codifies ways through which legal legal systems treat violations—typically, but not exclusively, through countermeasures such as obligations imposing various types of sanction.

For the sake of illustration, consider this example, where a contrary-to-duty obligation prescribes a sanction.

Example 5.3 (National Consumer Credit Protection Act 2009)**.** *Section 29 (Prohibition on engaging in credit activities without a licence) of the act recites:*

> *(1) A person must not engage in a credit activity if the person does not hold a licence authorising the person to engage in the credit activity.*
> *Civil penalty: 2,000 penalty units.*

[21] [Stolpe, 2010b] proposed two definitions. Here, we report on the simpler one.

> *[. . .]*
> *Criminal penalty: 200 penalty units, or 2 years imprisonment, or both.*

We can read this as that it is prohibited to engage in credit activities without a licence, but if one does not do so, then it is obligatory that is punished.

The deontic logic literature on CTD reasoning is immense: other chapters in this handbook deal with this topic. However, two fundamental mainstreams can be mentioned here as particularly interesting.

A first line of inquiry is mainly semantic-based. Moving from well-known studies on dyadic obligations, CTD reasoning is interpreted in settings with ideality or preference orderings on possible worlds or states [Hansson, 1969]. The value of this approach is that the semantic structures involved are quite flexible: depending on the properties of the preference or ideality relation, different deontic logics can be obtained. This semantic approach has been fruitfully renewed in the '90 for example by [Prakken and Sergot, 1996; van der Torre, 1997], and most recently by works such as [Hansen, 2005; van Benthem *et al.*, 2013], which have confirmed the vitality of this line of inquiry.

The second mainstream is mostly proof-theoretic, which is of interest here because some contributions in this context were driven—to some extent—by ideas from legal theory insofar as they

- clearly distinguished in the language and in the logic structures representing *norms* from those representing *obligations*, i.e., the consequences generated by norms,
- followed the slogan mentioned above "*no logic of norms without attention to the normative systems in which they occur*" [Makinson, 1999b]. Such a slogan is crucial in legal reasoning, as legal norms interplay in legal system.

Examples, among others, are various systems springing from Input/Output Logic [Makinson and van der Torre, 2000; Makinson and van der Torre, 2001] and the system proposed by [Governatori and Rotolo, 2006]. While Input/Output approach mainly works by imposing some constraints on the manipulation of conditional norms, [Governatori and Rotolo, 2006] is first of all based on the introduction of the new non-classical operator $\otimes$: the reading of an expression like $a \otimes b \otimes c$ is that a is primarily obligatory, but if this obligation is violated, the secondary obligation is b, and, if the secondary (CTD) obligation b is violated as

well, then c is obligatory. The approach in [Governatori and Rotolo, 2006] also introduced a non-classical consequence relation $\vdash$ to characterise normative conditionals generating obligations. Hence, an expression like

$$Invoice \vdash PayBy7days \otimes Pay5\%Interest \otimes Pay10\%Interest \quad (4)$$

can be intuitively viewed as a norm meaning the following:

1. if *Invoice* is the case, then *PayBy7days* is obligatory, but,
2. if *PayBy7days* is obligatory and $\neg$*PayBy7days* is the case, then *Pay5%Interest* is obligatory, but
3. if *Pay5%Interest* is obligatory and $\neg$*Pay5%Interest* is the case, then *Pay10%Interest* is obligatory.

In other words, one may see (4) as the merge of some interrelated conditional obligations, one the reparation of the violation of another.Their reciprocal interplay makes them interconnected so that they cannot be viewed anymore as independent obligations.

The logic for $\otimes$ is of interest in the law, because it is meant to specifically model CTDs as reparative obligations also know as compensatory obligations [Governatori, 2015]. Indeed, this is a mechanism very much used in the law, where the violations of norms trigger other norms prescribing sanctions or leading to restorative effects (see, e.g., [Hart, 1994; Kelsen, 1967]). We offer some details about developments on $\otimes$ in the remainder of this subsection.

5.3.1 The $\otimes$ logic and its semantics

The language of $\otimes$-logic is based on the introduction in Classical Propositional Logic of the usual unary operators O and P and the set of n-ary operators $\otimes^n$ for $n \in \mathbb{N}^+$. There is no technical difficulty in $\otimes$ not being a binary operator: the reason why we define it in terms of a set of n-ary ones is mainly conceptual and is meant to exclude the nesting of $\otimes$–expressions. Consider $a \otimes \neg(b \otimes c) \otimes d$. The expression $\neg(b \otimes c)$ means either that b is not obligatory or that it is so but c does not compensate the violation of $\mathsf{O}b$. What does it mean this as a compensation of the violation of $\mathsf{O}a$? Also, what is the meaning of $a \otimes (b \oplus c) \otimes d$?

This section summarises the contribution developed by [Governatori *et al.*, 2016b; Governatori *et al.*, 2016a; Calardo *et al.*, 2018], which further develops ideas from [Governatori and Rotolo, 2006] by considering the interplay between $\otimes$ and classical propositional logic (in Hilbert-style systems).

Let us examine some axiom schemata and inference rules.

The first principle is the well-known one of syntax independence or, in other terms, that the deontic operators are closed under logical equivalence:

$$\frac{\bigwedge_{i=1}^{n}(a_i \equiv b_i)}{\bigotimes_{i=1}^{n} a_i \equiv \bigotimes_{i=1}^{n} b_i}\otimes\text{-RE}$$

Consider $a \otimes b \otimes a \otimes c$. The meaning of this chain is that a is obligatory, but if a is violated (meaning that $\neg a$ holds) then b is obligatory. If also b is violated, then a becomes obligatory. But we already know that we will incur in the violation of it, since $\neg a$ holds. Accordingly, we have the obligation of c. However, this is the meaning of the $\otimes$-chain: $a \otimes b \otimes c$.

The above example shows that duplications of formulae in $\otimes$-chains do not contribute to the meaning of the chains themselves. This is a reason to adopt the following axioms to remove (resp., introduce) an element from (to) a chain if an equivalent formula occurs on the left of it.

$$\bigotimes_{i=1}^{n} a_i \equiv \bigotimes_{i=1}^{k-1} a_i \otimes \bigotimes_{i=k+1}^{n} a_i \text{ where } a_j \equiv a_k,\ j < k \qquad (\otimes\textbf{-contraction})$$

The above axiom and inference rule correspond to the minimal $\otimes$-logic $\mathsf{E}^{\otimes}$.

The next axiom provides a consistency principle for $\otimes$-chains. Given that we use classical propositional logic as the underlying logic, it is not possible that an $\otimes$-chain and its negation hold at the same time. What about when $\otimes$-chains like $a \otimes b \otimes c$ and $\neg(a \otimes b)$ hold. The first chain states that a is obligatory and its violation is compensated by b, which in turn is itself obligatory and it is compensated by c. The second expression states that 'either it is not the case that a is obligatory, but if it is so, then its violation is not compensated by b'. Accordingly, the combination of the two expressions should result in a contradiction. To ensure this, we must assume the following axioms that allow us to derive, given a chain, all its sub-chains with the same initial element(s).

$$a_1 \otimes \cdots \otimes a_n \rightarrow a_1 \otimes \cdots \otimes a_{n-1},\ n \geq 2 \qquad (\otimes\textbf{-shortening})$$

As expected, $\otimes$-chains are meant to generate obligations. In general:

$$a_1 \otimes \cdots \otimes a_n \wedge \bigwedge_{i=1}^{k<n} \neg a_i \rightarrow \mathsf{O}a_{k+1} \qquad (\mathsf{O}\textbf{-detachment})$$

In the simplest case, this amounts to $a_1 \otimes \cdots \otimes a_n \rightarrow \mathsf{O}a_1$: if, for example, the negation of the first element does not hold, we can infer the

obligation of the second element. A possible intuition behind schema O-**detachment** is that it can be used to determine which are the obligations that can be complied with. For example, since $\neg a_1$ holds, then we know that it is no longer possible to comply with the obligation of a_1. In a similar way, we could ask what are the parts of norms which are effective in a particular situation. In this case, instead of detaching an obligation we could detach an $\otimes$-chain. Accordingly, we formulate the following axiom:

$$a_1 \otimes \cdots \otimes a_n \wedge \neg a_1 \rightarrow a_2 \otimes \cdots \otimes a_n \qquad (\otimes\textbf{-detachment})$$

where $a_2 \otimes \cdots \otimes a_n$ does non contain a_1 or formulae equivalent to it.

Notice that the above detachment axioms do not explicitly mention that the negations of the first k elements of an $\otimes$-chain are violations. These axioms address this aspect:

$$a_1 \otimes \cdots \otimes a_n \wedge \bigwedge_{i=1}^{k<n} (\mathsf{O}a_i \wedge \neg a_i) \rightarrow \mathsf{O}a_{k+1} \quad (\mathsf{O}\textbf{-violation-detachment})$$

$$a_1 \otimes \cdots \otimes a_n \wedge \mathsf{O}a_1 \wedge \neg a_1 \rightarrow a_2 \otimes \cdots \otimes a_n \qquad (\otimes\textbf{-violation-detachment})$$

The proposed semantics for the $\otimes$-logic is called *sequence semantics*, which is is an extension of neighbourhood semantics. The extension is twofold: (1) a second neighbourhood-like function is introduced, and (2) the new function generates a set of sequences of sets of possible worlds instead of set of sets of possible worlds.

Definition 5.1. *A* sequence frame *is a structure* $\mathcal{F} = \langle W, \mathcal{C}, \mathcal{N} \rangle$*, where*

- W *is a non empty set of possible worlds,*
- $\mathcal{C}$ *is a function with signature* $W \rightarrow 2^{(2^W)^n}$ *such that for every world* w*, every* $X \in \mathcal{C}_w$ *is closed under s-zipping*[22]*.*
- $\mathcal{N}$ *is a function with signature* $W \rightarrow 2^{2^W}$*.*

Definition 5.2. *A* sequence model *is a structure* $\mathcal{M} = \langle \mathcal{F}, V \rangle$*, where*

- $\mathcal{F}$ *is a sequence frame, and*
- V *is a valuation function,* V*:* *Prop* $\rightarrow 2^W$*.*

$\otimes$- expressions are evaluated as follows:

[22]The operation of *s-zipping* corresponds to the removal of repetitions or redundancies occurring in sequences of sets of worlds [Governatori *et al.*, 2016b; Governatori *et al.*, 2016a; Calardo *et al.*, 2018]. It is required to capture the intuition described for the $\otimes$-shortening axioms.

Definition 5.3. *The valuation function for a sequence model is a follows:*

- *usual for atoms and boolean conditions,*
- $w \models \otimes_{i=1}^{n} a_i$ *iff* $\langle \|a_1\|_V, \ldots, \|a_n\|_V \rangle \in \mathcal{C}_w$,
- $w \models \mathsf{O}a$ *iff* $\|a\|_V \in \mathcal{N}_w$.

Various logical systems exist—based on the mentioned schemata and semantics—and the corresponding soundness and completeness results of these logics have been proved by [Governatori *et al.*, 2016b; Governatori *et al.*, 2016a; Calardo *et al.*, 2018].

5.4 Normative dynamics in the law

One peculiar feature of many normative systems, such as the law, is that it necessarily takes the form of a dynamic normative system [Kelsen, 1991; Hart, 1994]. Despite the importance of norm-change mechanisms, the logical investigation of legal dynamics is still relatively underdeveloped. However, recent contributions exist and this section is devoted to a brief sketch of this rapidly evolving literature.

Alchourrón and Makinson were the first to logically study the changes of a *legal code* [Alchourrón and Makinson, 1981; Alchourrón and Makinson, 1982; Alchourrón and Bulygin, 1981]. The addition of a new norm n causes an enlargement of the code, consisting of the new norm plus all the regulations that can be derived from n. Alchourrón and Makinson distinguish two other types of change. When the new norm is incoherent with the existing ones, we have an *amendment* of the code: in order to coherently add the new regulation n, we need to reject those norms that conflict with n. Finally, *derogation* is the elimination of a norm n together with whatever part of the legal code that implies n.

Alchourrón, Gärdenfors and Makinson [1985] inspired by the works above proposed the so called general AGM framework for belief revision. This area proved to be a very fertile one and the phenomenon of revision of logical theories has been thoroughly investigated. As is well-known, the AGM framework distinguishes three types of change operation over theories. Contraction is an operation that removes a specified sentence ϕ from a given theory Γ (a logically closed set of sentences) in such a way as Γ is set aside in favour of another theory Γ_{ϕ}^{-} which is a subset of Γ not containing ϕ. Expansion operation adds a given sentence ϕ to Γ so that the resulting theory Γ_{ϕ}^{+} is the smallest logically closed set that contains both Γ and ϕ. Revision operation adds ϕ to Γ but it is ensured that the resulting theory Γ_{ϕ}^{*} be consistent [Alchourrón *et al.*, 1985]. Alchourrón, Gärdenfors and Makinson argued that, when Γ is a code of legal norms,

contraction corresponds to norm derogation (norm removal) and revision to norm amendment.

It is then natural to ask if belief revision offers a satisfactory framework for the problem of norm revision in the law. Some of the AGM axioms seem to be rational requirements in a legal context, whereas they have been criticised when imposed on belief change operators. An example is the *success* postulate, requiring that a new input must always be in the belief set. It is reasonable to impose such a requirement when we wish to enforce a new norm or obligation. However, it gives rise to irrational behaviours when imposed to a belief set, as observed in [Gabbay *et al.*, 2003].

The AGM operation of contraction is perhaps the most controversial one, due to some postulates such as recovery [Governatori and Rotolo, 2010; Wheeler and Alberti, 2011], and to elusive nature of legal changes such as derogations and repeals, which are all meant to contract legal effects but in remarkably different ways [Governatori and Rotolo, 2010]. Standard AGM framework is of little help here: it has the advantage of being very abstract—it works with theories consisting of simple logical assertions—but precisely for this reason it is more suitable to capture the dynamics of obligations and permissions than the one of legal norms.

Difficulties behind AGM have been considered and some research has been carried out to reframe AGM ideas within reasonably richer rule-based logical systems able to capture the distinction between norms and legal effects [Stolpe, 2010c; Rotolo, 2010]. However, these attempts suffer from some drawbacks: they fail to handle reasoning on deontic effects and are based on a very simple representation of legal systems.

In fact, it is hard in AGM to represent how the same set of legal effects can be contracted in many different ways, depending on how norms are changed. These difficulties have been addressed in logical frameworks combining AGM ideas with richer rule-based logical systems, such as standard or Defeasible Logic [Rotolo, 2010; Governatori *et al.*, 2013b] or Input/Output Logic [Boella *et al.*, 2009; Stolpe, 2010c]. [Wheeler and Alberti, 2011] suggested a different route, i.e., employing in the law existing techniques—such as iterated belief change, two-dimensional belief change, belief bases, and weakened contraction—that can obviate problems identified in [Governatori and Rotolo, 2010] for standard AGM.

In general, any comprehensive logical model of norm change in the law has to take care of the following aspects:

1. the law usually regulates its own changes by setting specific norms whose peculiar objective is to change the system by stating what and how other existing norms should be modified;

2. since legal modifications are derived from these peculiar norms, they can be in conflict and so are defeasible;

3. legal norms are qualified by temporal properties, such as the time when the norm comes into existence and belongs to the legal system, the time when the norm is in force, the time when the norm produces legal effects, and the time when the normative effects hold.

To sum up, AGM-like frameworks have the advantage of being very abstract but works with theories consisting of simple logical assertions. For this reason, it is perhaps suitable to capture the dynamics of obligations and permissions, not of norms: the former ones are just possible effects of the application of norms and their dynamics do not necessarily require to remove or revise norms, but correspond in most cases to instances of the notion of *norm defeasibility* [Governatori and Rotolo, 2010] (see Section 6).

Hence, normative dynamics can be hardly modelled without considering temporal and defeasible reasoning. For this reason, previous works [Governatori *et al.*, 2005a; Governatori *et al.*, 2007b; Governatori and Rotolo, 2010] proposed to combine a rule-based system like Defeasible Logic with some forms of temporal reasoning. We will resume these ideas in Section 14.

6 Defeasibility in legal reasoning

One key idea of most logical accounts of the law is that legal reasoning is defeasible, namely, that we may have reasons to abandon certain legal conclusions even though there was no apparent mistake in previously supporting them [Sartor, 2005]. In legal theory, H.L.A. Hart was the first who illustrated this idea by saying, for instance, that "there are positive conditions required for the existence of a valid contract" but there are reasons that can defeat that existence claim, "even though all these conditions are satisfied" [Hart, 1951, p. 152]. The concept of defeasibility may have in the law different connotations.

6.1 Meanings of 'defeasibility' in the law

Consider art. 2051 of the Italian civil code: "A person is liable for damage caused by things in his custody except where he shows evidence of a fortuitous case". This legal provision states that the fault is not required to show the liability of the receiver for damage caused by things in safekeeping, thus highlighting the fact that the applicability conditions of

legal norms include both conditions that should be proved and conditions that should not be refuted (in this case, the fact that the receiver is at fault) [Sartor, 1995].

Conditions of the latter type can be explicit, like in the above provision, but are most often implicit. In general, the fact is that the statement of a norm can never mention all the relevant issues that might possibly be of relevance for its application, and in particular all its possible exceptions. This 'openness' to possible exceptions is a characteristic feature of legal norms and is known to be a peculiar aspect of legal *defeasibility*.

Defeasibility in legal norms breaks down, roughly, into the following issues:

Conflicts. Norms can conflict, namely, they may lead to incompatible legal effects. Conceptually, conflicts can be of different types, according to whether two conflicting norms

1. are such that one is an exception to the other (i.e., one is more specific than the other); this type of conflict can be solved using the principle *lex specialis*, which gives priority to the more specific norms (i.e., the exceptions);
2. have a different ranking status; this type of conflict can be solved using the principle *lex superior*, which gives priority to the norm from the higher authority;
3. have been enacted at different times; this type of conflict can be solved using the principle *lex posterior*, which gives priority to the norm enacted later.

Exclusionary norms. Some norms provide one way to explicitly undercut other norms, namely, to make them inapplicable. For example: "Art. 3 s not applicable in jurisdiction x".

Contributory reasons or factors. It is not always possible to formulate precise norms, even defeasible ones, for aggregating the factors relevant for resolving a legal issue. For example: "The educational value of a work needs to be taken into consideration when evaluating whether the work is covered by the copyright doctrine of fair use".

There are however more general reasons why legal reasoning should be viewed as defeasible. In fact, not all legal norms distinguish different types of applicability conditions (what should be proved and what should not be refuted), or not all norms admit exceptions or can be defeated. Independently of this, at a very general perspective, one may argue that legal reasoning is part of human cognition, the latter being inherently

defeasible [Pollock, 1995]; or, focussing more on the nature of law, that, even when norms seem to support indisputable conclusions, they are used in legal disputes or, more generally, in legal argumentative settings where arguments and counter-arguments dialectically interact.

When looking at the law through an argumentative lens, we may distinguish inference-based defeasibility, process-based defeasibility, and theory-based defeasibility [Prakken and Sartor, 2004].

Inference-based defeasibility covers the fact that legal conclusions, though correctly supported by certain pieces of information, cannot be derived when the knowledge base including this information is expanded with further pieces of information.

Process-based defeasibility addresses the dynamic aspects of defeasible reasoning. As for legal reasoning, a crucial observation here is that it often proceeds according to the norms of legal procedures, such as those regulating the allocation of the burden of proof.

Theory-based defeasibility regards the evaluation and the choice of theories which explain and systematize the available legal input information (such as a set of precedents): when a better theory becomes available, inferior theories are to be abandoned.

The remainder of this section briefly discusses aspects of the first two types of defeasibility. As for the the third type (theory-based defeasibility), the interested reader can still find a good primer in [Prakken and Sartor, 2004, sec. 4].

6.2 Defeasibility and argumentation layers in the law

Defeasible reasoning has been largely investigated in philosophy, logic, and AI by usually working on the concept of *inference-based defeasibility* [Makinson, 2005]. In this sense, defeasibility is formally interpreted within non-monotonic logics, namely, in logics whose underlying consequence relation does *not* enjoy the property of monotonicity, according to which conclusions never decrease if more knowledge is added. Since non-monotonicity means that a logic lacks a property, its positive interpretation is open to many options. In regard to modeling legal reasoning, since the Nineties the most preferred approach (especially in the AI&Law community) has been to develop argumentation systems (see, e.g., [Prakken and Sartor, 1996][23].

[23] Although it does not consider the most recent proposals, a still good introductory discussion can be found in [Prakken and Sartor, 2002]

However, other approaches in AI&Law have rather focused on process-based defeasibility [Gordon, 1995; Lodder, 1999; Prakken and Sartor, 1996; Bench-Capon *et al.*, 2000; Gordon *et al.*, 2007].

The advantage of all these approaches is that they intuitively capture the dialectal nature of legal reasoning by clearly considering its different layers. In particular, we need at least to distinguish a logical layer, a dialectical layer, and a procedural layer of legal arguments [Prakken and Sartor, 2002; Prakken and Vreeswijk, 2002].

Logical Layer The logical layer deals with the underlying language that is used to build legal arguments. Many languages and reasoning methods can be used for this purpose, such as deduction, induction, abduction, analogy, and case-based reasoning, provided that such methods are formalisable as logical systems[24]. If the underlying language refers to logic **L**, arguments can roughly correspond to proofs in **L** [Prakken and Sartor, 2002]. It may be argued that most (legal) argumentation systems are based on a *monotonic* consequence relation, since each single argument cannot be revised but can only be invalidated by *other* arguments (or better, counter-arguments) [Prakken and Vreeswijk, 2002]: it is the exchange of arguments and counter-arguments that makes the system non-monotonic. However, this is not strictly required: when the underlying logic is itself non-monotonic, an argumentation system can be simply seen as an alternative way to compute conclusions in that non-monotonic logic [Governatori *et al.*, 2004][25].

Suppose we resort to a rule-based logical system where rules have the form $\phi_1, \ldots, \phi_n \Rightarrow \phi$ and represent defeasible legal norms. An argument for a legal conclusion ϕ can typically have a tree-structure, where nodes correspond to literals and arcs correspond to the rules used to obtain these literals; hence, the root corresponds to ϕ, the leaf nodes to the primitive premises, and for every node corresponding to any literal ψ, if its children are $\psi_1, \ldots, \psi_n$, then there is a rule whose antecedents are these literals [Governatori *et al.*, 2004].

Argumentation systems, however, do not need in general to specify the internal structure of their arguments [Dung, 1995], so this assump-

[24]The application of these reasoning methods in the law have been studied by legal logicians, but space reasons prevent us to handle here this discussion. See [Sartor, 2005].

[25]If we embed within this language any deontic operators, we will obtain a way to deal with the defeasibility of the corresponding deontic concepts [Nute, 1998]. In general, various forms of interaction can be found among defeasibility, deontic concepts and normative systems. See [Sartor, 2005].

tion applies, too, to the legal domain. In this perspective, any (legal) argumentation system $\mathcal{A}$ is a structure $(A, \rightsquigarrow)$, where A is a non-empty set of arguments and $\rightsquigarrow$ is binary attack relation on A: for any pair or arguments a and b in A, $a \rightsquigarrow b$ means that a attacks b. This leads us to discuss the dialectical layer.

Dialectical Layer The dialectical layer addresses many interesting issues, such as when legal arguments conflict, how they can be compared and what legal arguments and conclusions can be justified.

Different types of attacks and defeat relations can apply to legal arguments. [Pollock, 1995]'s original distinction between *rebutting* and *undercutting* is almost universally accepted in the legal-argumentation literature [Prakken and Sartor, 2002; Prakken and Sartor, 2004]. An argument A_1 rebuts an argument A_2 when the conclusion of A_1 is equivalent to the negation of the conclusion of A_2. The rebutting relation is symmetric. For example, if arguments are built using rules representing legal norms (regulating, for example, smoking in public spaces), a conflict of this type at least corresponds to a clash between the conclusions obtained from two norms (for example, one prohibiting and another permitting to smoke). The undercutting is when an argument challenges a rule of inference of another argument. This attack relation is not symmetric and occurs when an argument A_1 supporting the conclusion ϕ has some ground ψ but another argument A_2 states that ψ is not a proper ground for ϕ. To put it very simple, if one builds an argument A_1 for ϕ using the rules $\Rightarrow \psi$ and $\psi \Rightarrow \phi$ but we contend that ψ does not support ϕ, then we undercut A_1.

Conflicts between legal arguments can be solved using specific legal-domain dependent priority criteria such as, as we said, *lex specialis*, *lex superior*, and *lex posterior*. However, such criteria can conflict, too, so some researchers argued that they must be defeasible [Prakken and Sartor, 1997; Prakken, 1995].

In general, assessing conflicting legal arguments cannot work if we only examine single pairs of arguments. In fact, we need to consider all the arguments to establish what legal conclusions win and are justified in a legal dispute. Argumentation theory usually distinguishes among *justified*, *defensible* and *overruled* arguments [Dung, 1995]. Justified arguments are those which basically survive from all attacks, the defensible ones leave the dispute undecided, and the overruled ones are those defeated by a justified argument [Prakken and Vreeswijk, 2002]. Doing so, we may have to capture interesting complex argumentative patterns. For

instance, consider this argumentation system:

$$(A = \{A_1, A_2, A_3\}, \rightsquigarrow = \{\langle A_1, A_2\rangle, \langle A_2, A_1\rangle, \langle A_3, A_2\rangle\})$$

The argument A_1 is attacked by the argument A_2 but it may be reinstated when a third argument A_3 attacking A_2 comes into play [Prakken and Vreeswijk, 2002]. This is an example of a reasoning pattern known in argumentation theory as *reinstatement*, which is relevant, for instance, in legal evidential reasoning: suppose Henry was killed yesterday and John was charged with that crime. Tom argues that John did not kill yesterday Henry, but Nino testifies that John indeed killed him. Tom original argument can be reinstated by another testimony showing that that Nino was drunk yesterday.

Another interesting pattern regards the so-called floating conclusions [Horty, 2002]. Consider the following two arguments (represented as chains of rules):

$$\begin{array}{llll} A_1 & \mathit{testimonyA} & \Rightarrow \mathit{JohnShootHenry} & \Rightarrow \mathit{guilty} \\ A_2 & \mathit{testimonyB} & \Rightarrow \mathit{JohnPoisonHenry} & \Rightarrow \mathit{guilty} \end{array} \tag{5}$$

The two arguments lead to the same conclusion but one sub-argument of A_1 attacks one sub-argument of A_2 and vice versa (the fact that John shot Henry excludes that John poisoned Henry and vice versa). One may say that John is anyway guilty, whatever argument we may prefer, but we can also argue that the two testimonies undermine each other, so no conclusion could be obtained.

Procedural Layer The procedural layer considers the ways through which conclusions are dynamically reached in legal disputes. Indeed, disputes can be reconstructed in the form of dialogues, namely of players' dialectical moves [Gordon, 1995; Prakken, 2001]. Legal disputes are regulated by procedural rules stating what dialogue moves (claiming, challenging, conceding, etc.) are possible, when they are legal, what effects the players get from them, and under what conditions a dispute terminates [Gordon, 1995; Lodder, 1999] (in general, see [Walton and Krabbe, 1995]).[26]

A basic and fundamental question of the procedural layer regards how to govern and allocate the burden of proof [Prakken, 2001]. For example, basic dialogue protocols of 2-player civil disputes are defined

[26]The idea that justice depends on formal procedures governing public deliberation and dialogues has been defended, among others, in [Rescher, 1977; Rawls, 1971] and in the law in [Alexy, 1989].

on account of the requirement that the plaintiff begins the dispute with his claim and has to propose, to win, at least one justified argument which support such a claim. The burden of the defendant is not in principle the same, as it may be sufficient in most cases for her to oppose the plaintiff argument moves with merely defensible counter-arguments. The concept of legal burden of proof is very complex and its logical treatment is difficult: the interested reader can refer to [Prakken and Sartor, 2009]. Even more complex is to handle the interplay between the dialectical and the procedural layers [Prakken, 2001]. To appreciate this, consider the example on floating conclusions of Formula 5. Here, players, if dynamically modeled at the procedural layer would certainly postpone their judgment and subsequently challenge both the testimonies and test their credibility [Prakken and Sartor, 2004].

Part III
A Logic of Norms and Normative Systems – A Unifying Framework for the Law

7 Introduction

In this part we are going to provide the foundations of a general framework for the logic of legal norms and legal systems. The unifying framework for that purpose is a family of variants Defeasible Logic [Nute, 1994]. Defeasible Logic is a simple, flexible, extensible non-monotonic formalism, and has been used in the literature to cover all main aspects of the logic of norms, as applied in the law.

In particular, we will show how different variants of Defeasible Logic can provide the solutions to all the main issues mentioned in Part II of this chapter, i.e.:

- a simple but plausible notion of legal defeasibility;
- a logic covering several types of norms and normative judgements;
- a representation of legal systems;
- ways for handling contrary-to-duty reasoning;
- a comprehensive model of norm change in the law;

- a deontic logic for modelling legal interpretation which is based on the distinction between legal provision and legal norm.

For space reasons, the purpose of this part of the chapter is not to offer a comprehensive technical presentation of all variants of Defeasible Logic for modelling legal reasoning. Rather we would like to provide the reader with the main conceptual and logical intuitions behind each formalism, sometimes referring to some technical results when this is relevant for modelling deontic reasoning in the law.

We will begin by presenting the basic variants of Defeasible Logic.

8 Basic defeasible logic

Let us define a set Lit of literals $\{a, b, c, \dots\}$, such that, if a is a literal, $\sim a$ denotes the complementary literal (if a is a positive literal p then $\sim a$ is $\neg p$; and if a is $\neg p$, then $\sim a$ is p).

Knowledge in Defeasible Logic is structured in three components:

- A set of facts (corresponding to indisputable statements represented as literals, where a literal is either an atomic proposition or its negation).
- A set of rules. A rule establishes a connection between a set of premises and a conclusion. In particular, for reasoning with norms, it is reasonable to assume that a rule provides the formal representation of a norm. Accordingly, the premises encode the conditions under which the norm is applicable, and the conclusion is the normative effect of the norm.
- A preference relation over the rules. The preference relation just gives the relative strength of rules. It is used in contexts where two rules with opposite conclusions fire simultaneously, and determines that one rule overrides the other in that particular context.

Formally, the knowledge in the logic is organised in Defeasible Theories, where a Defeasible Theory D is a structure

$$(F, R, \prec) \tag{6}$$

where F is the set of facts, R is the set of rules, and $\prec$ is a binary relation over the set of rules, i.e., $\prec \subseteq R \times R$.[27]

[27] Defeasible Logic does not impose any property for $\prec$. However, in many applications it is useful to assume that the transitive closure to be acyclic to prevent situations where, at the same time a rule overrules another rule and it is overridden by it.

A rule is formally a binary relation between, a set premises and a conclusion. Thus if Lit is the set of literals, the set Rule of all rules is:

$$\text{Rule} \subseteq 2^{\text{Lit}} \times \text{Lit}. \tag{7}$$

Accordingly, a rule is an expression with the following form:[28]

$$r\colon a_1, \ldots, a_n \hookrightarrow c \tag{8}$$

where r is a unique label identifying the rule. Given that a rule is a relation, we can ask what is the strength of the link between the premises and the conclusion. We can distinguish three different strengths: (i) given the premises the conclusion always holds, (ii) given the premises the conclusion holds sometimes, and (iii) given the premises the opposite of the conclusions does not hold. Therefore, to capture theses types Defeasible Logic is equipped with three types of rules: *strict rules*, *defeasible rules* and *defeaters*. We will use $\rightarrow$, $\Rightarrow$ and $\rightsquigarrow$ instead of $\hookrightarrow$ to represent, respectively, strict rules, defeasible rules and defeaters. We will continue to use $\hookrightarrow$ for a rule when the strength is either not known or irrelevant.

Given a rule like rule r in (8) we use the following notation to refer to the various elements of the rule. $A(r)$ denotes the *antecedent* or *premises* of the rule, in this case, $\{a_1, \ldots, a_n\}$, and $C(r)$ denotes the *conclusion* or *consequent*, that is, c. From time to time we use *head* and *body* of a rule to refer, respectively, to the consequent and to the antecedent of the rule.

Strict rules are rules in the classic sense: whenever the premises are indisputable so is the conclusion. Strict rules can be used to model legal definitions that do not admit exceptions, for example the definition of minor: "'minor' means any person under the age of eighteen years". This definition can be represented as

$$age(x) < 18yrs \rightarrow minor(x). \tag{9}$$

Defeasible Rules are rules such that the conclusions normally or typically follows from the premises, unless there are evidence or reasons to the contrary.

Defeaters are rules that do not support directly the derivation of a conclusion, but that can be used to prevent a conclusion.

We illustrate defeasible rules and defeaters with the help of the definition of complaint from the Australian Telecommunication Consumer Protections Code 2012 TCP-C268_2012 May 2012 (TCPC).

[28] More correctly, we should use $r\colon \{a_1, \ldots, a_n\} \hookrightarrow c$. However, to improve readability, we drop the set notation for the antecedent of rule.

> **Complaint** means an expression of dissatisfaction made to a Supplier in relation to its Telecommunications Products or the complaints handling process itself, where a response or Resolution is explicitly or implicitly expected by the Consumer.
>
> An initial call to a provider to request a service or information or to request support is not necessarily a Complaint. An initial call to report a fault or service difficulty is not a Complaint. However, if a Customer advises that they want this initial call treated as a Complaint, the Supplier will also treat this initial call as a Complaint.
>
> If a Supplier is uncertain, a Supplier must ask a Customer if they wish to make a Complaint and must rely on the Customer's response.

Here is a (simplified) formal representation:

$$tcpc_1 : \mathit{ExpressionDissatisfaction} \Rightarrow \mathit{Complaint}$$
$$tcpc_2 : \mathit{InformationCall} \Rightarrow \neg \mathit{Complaint}$$
$$tcpc_3 : \mathit{ProblemCall}, \mathit{FirstCall} \rightsquigarrow \mathit{Complaint}$$
$$tcpc_4 : \mathit{AdviseComplaint} \Rightarrow \mathit{Complaint}$$

where $tcpc_1 \prec tcpc_2$ and $tcpc_2 \prec tcpc_4$.

The first rule $tcpc_1$ sets the basic conditions for something to be a complaint. On the other hand, rule $tcpc_2$ provides an exception to the first rule, and rule $tcpc_4$ is an exception to the exception provided by rule $tcpc_2$. Finally, $tcpc_3$ does not alone warrant the call to be a complaint (though, it does not preclude the possibility that the call turns out to be a complaint; hence the use of a defeater to capture this case).

Defeasible Logic is a constructive logic. This means that at the heart of it we have its proof theory, and for every conclusion we draw from a defeasible theory we can provide a proof for it, giving the steps used to reach the conclusion, and at the same time, providing a (formal) explanation or justification of the conclusion. Furthermore, the logic distinguishes *positive* and *negative* conclusion, and the strength of a conclusion. This is achieved by labelling each step in a derivation with a proof tag. As usual a derivation is a (finite) sequence of formulas, each obtained from the previous ones using inference conditions.

Let D be a Defeasible Theory. The following are the proof tags we consider for basic Defeasible Logic:

$+\Delta$ if a literal p is tagged by $+\Delta$, then this means that p is provable using only the facts and strict rules in a defeasible theory. We also say that p is *definitely provable* from D.

$-\Delta$ if a literal p is tagged by $-\Delta$, then this means that p is refuted using only the facts and strict rules in a defeasible theory. In other terms, it indicates that the literal p cannot be proved from D using only facts and strict rules. We also say that p is *definitely refuted* from D.

$+\partial$ if a literal p is tagged by $+\partial$, then this means that p is *defeasibly provable* from D.

$-\partial$ if a literal p is tagged by $-\partial$, then this means that p is *defeasibly refutable* from D.

Some more notation is needed before explaining how tagged conclusions can be asserted. Given a set of rules R, we use R_x to indicate particular subsets of rules: R_s for strict rules, R_d for defeasible rules, R_{sd} for strict or defeasible rules, R_{dft} for defeaters; finally $R[q]$ denotes the rules in R whose conclusion is q.

Provability is based on the concept of a *derivation* (or proof) in a theory D. A derivation is a finite sequence $P = (P(1), \ldots, P(n))$ of tagged literals satisfying four conditions (which correspond to inference rules for each of the four kinds of conclusion). $P(1..i)$ denotes the initial part of the sequence P of length i.

There are two ways to prove $+\Delta p$ at the n-th step of a derivation: the first is that p is one of the facts of the theory. The second case is when we have a strict rule r for p and all elements in the antecedent of r have been definitely proved at previous steps of the derivation.

For $-\Delta p$ we have to argue that there is no possible way to derive p using facts and strict rules. Accordingly, p must not be one of the facts of the theory, and second for every rule in $R_s[p]$ (all strict rules which are able to conclude p) the rule cannot be applied, meaning that at least one of the elements in the antecedent of the rule has already refuted (definitely refuted). The base case is where the literal to be refuted is not a fact and there are no strict rules having the literal as their head.

Defeasible derivations have a three phases argumentation like structure[29]. To show that $+\partial p$ is provable (i.e., that p is defeasibly provable)

[29]The relationships between Defeasible Logic and argumentation are, in fact, deeper than the similarity of the argumentation like proof theory. [Governatori *et al.*, 2004] prove characterisation theorems for defeasible logic variants and Dung style argumentation semantics [Dung, 1995]. In addition [Governatori, 2011] proved that the Carneades argumentation framework [Gordon *et al.*, 2007], widely discussed in the

at step n of a derivation we have to:[30]

1. give an argument for p;
2. consider all counterarguments for p; and
3. rebut each counterargument by either:
   (a) showing that the counterargument is not valid;
   (b) providing a valid argument for p defeating the counterargument.

In this context, in the first phase, an argument is simply a strict or defeasible rule for the conclusion we want to prove, where all the elements are at least defeasibly provable. In the second phase we consider all rules for the opposite or complement of the conclusion to be proved. Here, an argument (counterargument) is not valid if the argument is not supported.[31] Here "supported" means that all the elements of the body are at least defeasibly provable.

Finally to defeasibly refute a literal, we have to show that either, the opposite is at least defeasible provable, or show that an exhaustive search for a constructive proof for the literal fails (i.e., there are rules for such a conclusion or all rules are either 'invalid' argument or they are not stronger than valid arguments for the opposite).

For the sake of illustration, for any literal p here below are the formal proof conditions for $+\partial p$ and $-\partial p$[32]:

AI and Law literature, turns out to be just a syntactic variant of Defeasible Logic.

[30]Here we concentrate on proper defeasible derivations. In addition we notice that a defeasible derivations inherit from definite derivations, thus we can assert $+\partial p$ if we have already established $+\Delta p$.

[31]It is possible to give different definition of support to obtain variants of the logic tailored for various intuitions of non-monotonic reasoning. [Billington *et al.*, 2010] show how to modify the notion of support to obtain variants capturing such intuitions, for example by weakening the requirements for a rule to be supported: instead of being defeasibly provable a rule is supported if it is possible to build a reasoning chain from the facts ignoring rules for the complements.

[32]Notice that defeaters are only considered in (2.3) when attacks to r are handled. In the other cases, when rules are expected to prove literals (or to reinstate them), this version of Defeasible Logic uses only strict and defeasible rules.

$+\partial$: If $P(i+1) = +\partial p$ then either
(1) $+\Delta p \in P(1..i)$ or
(2.1) $\exists r \in R_{sd}[p] \forall a \in A(r) : +\partial a \in P(1..i)$ and
(2.2) $-\Delta{\sim}p \in P(1..i)$ and
(2.3) $\forall s \in R[{\sim}p]$ either
  (2.3.1) $\exists a \in A(s) : -\partial a \in P(1..i)$ or
  (2.3.2) $\exists t \in R_{sd}[p]$ such that
    $\forall a \in A(t) : +\partial a \in P(1..i)$ and $t \prec s$.

$-\partial$: If $P(i+1) = -\partial p$ then
(1) $-\Delta p \in P(1..i)$ and
(2.1) $\forall r \in R_{sd}[B]\ \exists a \in A(r) : -\partial a \in P(1..i)$ or
(2.2) $+\Delta{\sim}p \in P(1..i)$ or
(2.3) $\exists s \in R[{\sim}p]$ such that
  (2.3.1) $\forall a \in A(s) : +\partial a \in P(1..i)$ and
  (2.3.2) $\forall t \in R_{sd}[p]$ either
    $\exists a \in A(t) : -\partial a \in P(1..i)$ or $t \not\prec s$.

Consider again the set of rules encoding the TCPC 2012 definition of complaint. Assume to have a situation where there is an initial call from a customer who is dissatisfied with some aspects of the service received so far where she asks for some information about the service. In this case rules $tcpc_1$ and $tcpc_2$ are both applicable (we assume that the facts of the case include the union of the premises of the two rules, but *AdviseComplaint* is not a fact). Here, $tcpc_2$ defeats $tcpc_1$, and $tcpc_4$ cannot be used. Hence, we can conclude $-\partial$*AdviseComplaint* and consequently $+\partial\neg$*Complaint* and $-\partial$*Complaint*. However, if the customer stated that she wanted to complain for the service, then the fact *AdviseComplaint* would appears in the facts. Therefore we can conclude $+\partial$*AdviseComplaint*, making then rule $tcpc_4$ applicable, and we can reverse the conclusions, namely: $+\partial$*Complaint* and $-\partial\neg$*Complaint*.

8.1 Further readings

For an in-depth presentation of propositional Defeasible Deontic Logic and its properties we refer the reader to [Antoniou *et al.*, 2001; Governatori *et al.*, 2004].

While the propositional Defeasible Logic we outlined above—and its variants—are able to model different features of legal reasoning (e.g., burden of proof [Governatori and Sartor, 2010] and proof standards [Governatori, 2011] covering and extending the proof standards discussed in [Gordon *et al.*, 2007]), some important characteristics of legal reasoning are missing, all of them being modelled in Part II of this chapter and

mentioned at the beginning of this part.

9 Defeasible deontic logic

Norms in a normative system can have (among others, but typically) the following functions:

1. to define the terms and concepts used in the normative system, and
2. to prescribe the behaviours the subjects of the normative system are meant to comply with.

The distinction just introduced is that of *constitutive rules* and *prescriptive rules* (see Section 4.2). The "mode" of the behaviours prescribed by the prescriptive rules is determined by deontic modalities (e.g., obligation, prohibition, permission). The Defeasible Logic presented in the previous section accounts for constitutive rules. To model prescriptive rules we have (i) to extend the language with deontic operators (ii) to use again the idea that rules are just binary relations and add a dimension, that is the *mode*, in the classification of rules. Hence rules can be classified according to their strength as well as their mode.

We concentrate now on the following deontic operators: O, P and F, respectively for obligation, permission and prohibition. In the language of Defeasible Deontic Logic the set of literals Lit is partitioned in *plain literals* and *deontic literals*. A plain literal is a literal in the sense of basic defeasible logic, while a deontic literal is obtained by placing a plain literal in the scope of a deontic operator or a negated deontic operator. Accordingly, expressions like $\mathsf{O}l$, $\neg\mathsf{P}l$ and $\mathsf{F}\neg l$ are deontic literals, where l is plain literal.

In Defeasible Deontic Logic rules are defined with the following signature

$$\text{Rule}\colon 2^{\text{Lit}} \times \text{PlainLit} \tag{10}$$

where PlainLit is the set of all plain literals. This means that the antecedent of a rule can contain both plain and deontic literal, but in any case the conclusion is plain literal. Thus the question is if the conclusions of rules are plain literals, where do we get deontic literals? The answer is that we have two different modes of for the rules. The first mode is that of constitutive rule, where the conclusion is an assert with the same mode as it appears in the rule (i.e., as an institutional fact); the second mode is that of prescriptive rule, where the conclusion is asserted with a deontic mode (where the deontic mode corresponds to one of the deontic

operators). Accordingly, a Defeasible Deontic Theory is a structure

$$(F, R^C, R^{\mathsf{O}}, \prec) \tag{11}$$

where R^C is a set of constitutive rules, and R^{O} is a set of prescriptive rules. Constitutive rules behaves as the rules in Basic Defeasible Logic, and we continue to use $\hookrightarrow$ to denote the arrow of a constitutive rule. $\hookrightarrow_{\mathsf{O}}$ for the arrow of a prescriptive rule.

The main idea is that given the constitutive defeasible rule

$$a_1, \ldots, a_n \Rightarrow_C b \tag{12}$$

we can assert b, given $a_1, \ldots, a_n$, thus the behaviour of constitutive rule is just the normal behaviour of rules we examined in the previous section. For prescriptive rules the behaviour is a different. From the rule

$$a_1, \ldots, a_n \Rightarrow_{\mathsf{O}} b \tag{13}$$

we conclude $\mathsf{O}b$ when we have $a_1, \ldots, a_n$. Thus we conclude the obligation of the conclusion of the rule, not just the conclusion of the rule.

The reasoning mechanism is essentially the same as that of basic defeasible presented in Section 8 with two differences. First, an argument can only be attacked by an argument of the same type. Thus if we have an argument consisting of a constitutive rule for p, a counterargument should be a constitutive rule for $\sim p$. The same applies for prescriptive rules. An exception to this is when we have a constitutive rule for p such that all its premises are provable as obligations. In this case the constitutive rule behaves like a prescriptive rule, and can be use as a counterargument for a prescriptive rule for $\sim p$, or the other way around. The second difference is that now the proof tags are labelled with either C, e.g., $+\partial_C p$, (for constitutive conclusions) or with O, e.g., $-\partial_{\mathsf{O}} q$ (for prescriptive conclusions). Accordingly, when we are able to derive $+\partial_{\mathsf{O}} p$ we can say that $\mathsf{O}p$ is provable.

This feature poses the question of how we model the other deontic operators (i.e., permission and prohibition). As customary in Deontic Logic, we assume the following principles governing the interactions of the deontic operators.[33]

$$\mathsf{O}{\sim}l \equiv \mathsf{F}l \tag{14}$$

$$\mathsf{O}l \wedge \mathsf{O}{\sim}l \rightarrow \bot \tag{15}$$

$$\mathsf{O}l \wedge \mathsf{P}{\sim}l \rightarrow \bot \tag{16}$$

[33] In the three formulas below $\rightarrow$ is the material implication of classical logic.

Principle (14) provides the equivalence of a prohibition with a negative obligation (i.e., obligation not). The second and the third are rationality postulates stipulating that it is not possible to have that something and its opposite are at the same time obligatory (15) and that a normative system makes something obligatory and its opposite is permitted (16). (14) gives us the immediate answer on how prohibition is modeled. A rule giving a prohibition can be modelled just as a prescriptive rule for a negated literal. This means that to conclude $\mathsf{F}p$ we have to derive $+\partial_{\mathsf{O}}\neg p$.

Consider Section 40 of the Australian Road Rules (ARR)[34]

> **Making a U–turn at an intersection with traffic lights**
> A driver must not make a U–turn at an intersection with traffic lights unless there is a U–turn permitted sign at the intersection.

The prohibition of making U-turns at traffic lights can be encoded by the following rule:

$$arr_{40a}\colon AtTrafficLigths \Rightarrow_{\mathsf{O}} \neg Uturn.$$

In a situation where *AtTrafficLights* is given we derive $+\partial_{\mathsf{O}}\neg Uturn$ which corresponds to $\mathsf{F}\,Uturn$.

Details of the proof theory are omitted (see [Governatori *et al.*, 2013a]). We just notice that the basic conditions presented earlier for the propositional case must handle the following intuitions:

- conflicts between rules can only occur between any pair constitutive rules like $a_1, \ldots, a_n \Rightarrow_C l$ and $b_1, \ldots, b_n \Rightarrow_C \sim l$, on the one side, and any pair of obligation rules such as $a_1, \ldots, a_n \Rightarrow_{\mathsf{O}} l$ and $b_1, \ldots, b_n \Rightarrow_{\mathsf{O}} \sim l$, or $a_1, \ldots, a_n \Rightarrow_{\mathsf{O}} l$ and $b_1, \ldots, b_n \rightsquigarrow_{\mathsf{O}} \sim l$;
- as we have alluded to, the proof theory works in such a way that the provability of literal l with mode O (i.e., $+\partial_{\mathsf{O}} l$) leads to deriving $\mathsf{O}l$, which means that through this conclusion we fire another rule where $\mathsf{O}l$ occurs in the antecedent;
- when $\mathsf{O}l$ is defeasibly disproved (by showing that $-\partial_{\mathsf{O}} l$), then $\mathsf{P}{\sim}l$ is positively proved.

[34] This norm makes use of "must not", to see that "must not" is understood as prohibition in legal documents see, the Australian National Consumer Credit Protection Act 2009, Section 29, whose heading is "Prohibition on engaging in credit activities without a licence", recites "(1) A person must not engage in a credit activity if the person does not hold a licence authorising the person to engage in the credit activity".

Under the above intuitions, proof theory for $\Rightarrow_C$ is standard propositional Defeasible Logic. As regards obligations, to show that any literal l is defeasibly provable as an obligation, there are two ways: (1) the obligation of l is a fact, or (2) l must be derived by the obligation rules of the theory. In the second case, three conditions must hold: (2.1) l does not appear as not obligatory as a fact, and $\sim l$ is not provable as an obligation using the set of modal facts at hand; (2.2) there must be a rule introducing the obligation for l which can apply; (2.3) every rule s for $\sim l$ is either discarded or defeated by a stronger rule for l.

The pending issue is how to model various types of permission. As we have already recalled above, two types of permissions have been discussed in literature following [von Wright, 1963] and [Alchourrón and Bulygin, 1984]: (i) weak permission, meaning that there is no obligation to the contrary; and (ii) strong permission, a permission explicitly derogates an obligation to the contrary. In this case we have an exception. For both types of permission we have that the obligation to the contrary does not hold. Defeasible Deontic Logic is capable to handle the two types of permission if we establish that $\mathsf{P}p$ is captured by $-\partial_{\mathsf{O}}{\sim}p$. The meaning of $-\partial_{\mathsf{O}}p$ is that p is refuted as obligation, or that it is not possible to prove p as an obligation; hence it means that we cannot establish that p is obligatory, thus there is no obligation contrary to $\sim p$.

We will return on this in Section 10.

9.1 Further readings

For an in-depth presentation of Defeasible Deontic Logic, its properties and a detailed analysis of how to use it to model obligations and permissions (and several ways to do it) we refer the reader to [Governatori *et al.*, 2013a]. The reader can also consult [Governatori *et al.*, 2005b] where some advanced extensions with time are analysed (on time in deontic reasoning see Section 13). Other applications of modal extensions of Defeasible Logic where deontic notions play a significant role are [Governatori and Rotolo, 2008a; Governatori and Rotolo, 2008b].

The reader may notice that modalities here and in subsequent Sections of Part III of this chapter are constructively defined by the proof conditions. However, this does not mean that they cannot interpreted in a standard model-theoretic semantics: see [Governatori *et al.*, 2012] for the details, which however, offer a rather technical discussion. We can simply observe that most of the corresponding modal logics are weaker than **K**, i.e., they are non-normal.

10 Modelling permissions

10.1 Permissions and defeasibility

The above framework, though simple, allows us to express some basic types of permissions as well as illustrate interesting connections with the idea of defeasibility. Let us again consider the distinction between weak and strong permission.

Weak Permission A first way to define permissions in Defeasible Deontic Logic consists in simply considering weak permissions and stating that the opposite of what is permitted is not provable as obligatory. Let us consider a normative system consisting of the following two rules:

$$\begin{aligned} r_1 &: \quad \mathit{Park}, \mathit{Vehicle} \Rightarrow_{\mathsf{O}} \neg \mathit{Enter} \\ r_2 &: \quad \mathit{Park}, \mathit{Emergency} \Rightarrow_{\mathsf{O}} \mathit{Enter}. \end{aligned}$$

Here, the normative system does not contain any permissive norm. However, since Defeasible Deontic Logic is a sceptical non-monotonic logic, in case both r_1 and r_2 fire we neither conclude that it is prohibited nor conclude that it is obligatory to enter, because we do not know which rule is stronger. Consequently, in this context, both $\neg \mathit{Enter}$ and Enter are weakly permitted.

As already argued, this is the most direct way to define the idea of weak permission: some q is permitted by a code iff q is not prohibited by that code. Accordingly, saying that any literal q is weakly permitted corresponds to the failure of deriving $\neg q$ using rules for O.

Explicit Permissions as Defeaters In Defeasible Deontic Logic any rule can be used to prevent the derivation of a conclusion. For instance, suppose there exists a norm that prohibits to U-turn at traffic lights unless there is a "U-turn permitted" sign:

$$\begin{aligned} r_1 &: \mathit{AtTrafficLight} \Rightarrow_{\mathsf{O}} \neg \mathit{Uturn} \\ r_2 &: \mathit{AtTrafficLight}, \mathit{UturnSign} \Rightarrow_{\mathsf{O}} \mathit{Uturn}. \end{aligned}$$

We use a defeasible rule for obligation to block the prohibition to U-turn. However, this is not satisfactory for a number of reasons: if we do not know whether r_2 is stronger than r_1, then the best we can say is that U-turn is weakly permitted. Furthermore, if r_2 prevails over r_1, we derive that U-turn is obligatory: indeed, r_2 does not express a permission (as suggested by saying that U-turn at traffic lights is permitted if there is a sign "U-turn permitted") but an obligation.

Thus, when permissions derogate to prohibitions, there are good reasons to argue that defeaters for O are suitable to express an idea of strong permission[35]. Explicit rules such as $r : a \rightsquigarrow_{\mathsf{O}} q$ state that a is a specific reason for blocking the derivation of $\mathsf{O}\neg q$ (but not for proving $\mathsf{O}q$), i.e., this rule does not support any conclusion, but states that $\neg q$ is deontically undesirable. Consider this example:

$$\begin{array}{ll} r_1 : & \mathit{Weekend}, \mathit{AirPollution} \Rightarrow_{\mathsf{O}} \neg \mathit{UseCar} \\ r_2 : & \mathit{Weekend}, \mathit{Emergency} \rightsquigarrow_{\mathsf{O}} \mathit{UseCar}. \end{array}$$

Rule r_1 states that on weekends it is forbidden to use private cars if a certain air pollution level is exceeded. Defeater r_2 is in fact an exception to r_1, and so it seems to capture the above idea that explicit permissive norms provide exceptions to obligations.

Explicit Permissions as Permissive Rules Another approach is based on introducing specific rules for deriving permissions [Makinson and van der Torre, 2003; Boella and van der Torre, 2003a]. Let us consider the following situation:

$$\begin{array}{ll} r_1 : & \mathit{Weekend}, \mathit{AirPollution} \Rightarrow_{\mathsf{O}} \neg \mathit{UseCar} \\ r_2' : & \mathit{Emergency} \Rightarrow_{\mathsf{P}} \mathit{UseCar}. \end{array}$$

As r_2 in the previous scenario, r_2' looks like an exception to r_1. The apparent difference between r_2 and r_2' is that the latter is directly used to prove that the use of the car is permitted ($\mathsf{P}\,\mathit{UseCar}$) in case of emergencies. Does it amount to a real difference?

Although r_2 is a defeater, it is specifically used to derive the strong permission to use the car, like r_2'. In addition, rules such as r_2' do not attack other permissive rules, but are in conflict only with rules for obligation intended to prove the opposite conclusion. This precisely holds for defeaters.

Moreover, let us suppose to have the defeater $s : a \rightsquigarrow_{\mathsf{P}} b$. Does s attack a rule like $\Rightarrow_{\mathsf{P}} \neg b$? If this is the case, s would be close to an obligation. The fact that $\mathsf{P}b$ does not attack $\mathsf{P}\neg b$ makes it pointless for s to introduce defeaters for P. But, if this is not the case, s could only attack $\Rightarrow_{\mathsf{O}} \neg b$, thus being equivalent to $s' : a \rightsquigarrow_{\mathsf{O}} b$. Although it is admissible to have defeaters, we do not need to distinguish defeaters for O from those for P.

[35] The idea of using defeaters to introduce permissions was introduced in [Governatori *et al.*, 2005b].

We see two arguments to differentiate the approaches that model permissions via $\rightsquigarrow_{\mathsf{O}}$ and via $\Rightarrow_{\mathsf{P}}$. The first argument is purely conceptual in that defeaters act only to block opposite conclusions. Hence, they are suitable for modelling only derogations but not permissive rights, which are incompatible with opposite obligations but are not primarily designed to be exceptions to prohibitions (see below, Example 10.2 for a concrete illustration). Another way to mark the difference between $\rightsquigarrow_{\mathsf{O}}$ and $\Rightarrow_{\mathsf{P}}$ is by stating that only the latter rule type admits ordered sequences of strong permissions in the head of a rule, which are either permissive rights or derogations to prohibitions. This second matter will be presented in the next subsection.

10.2 Permissions, obligations, and preferences

In Section 5.3.1 we have reported on a logic for the $\otimes$ operator, which was originally devised to model contrary-to-duty reasoning. [Governatori *et al.*, 2016a] introduced another non-classical operator $\odot$ to capture an analogous intuition for strong permission. As in the case of $\otimes$, given $a \odot b$, we can proceed through the $\odot$-chain to obtain the derivation of $\mathsf{P}b$. However, permissions cannot be violated, and consequently it does not make sense to obtain $\mathsf{P}b$ from $a \odot b$ and $\neg a$. In this case, the reason to proceed in the chain is rather that the normative system allows us to prove $\mathsf{O}\neg a$. Hence, $\odot$ still establishes a preference order among strong permissions and, in case the opposite obligation is in force, another permission holds.

The full logic is based on the following intuitions:

1. Permissive and obligation rules are represented as before by defeasible deontic rules. In particular, since the rule

$$Order \Rightarrow_{\mathsf{O}} Pay$$

   says that, if we send a purchase order, then we are defeasibly obliged to pay, analogously the rule

$$Order, Creditor \Rightarrow_{\mathsf{P}} \neg Pay$$

   states that if we send an order, in general we are not obliged to pay if we are creditors for the same amount.

2. Again, deontic rules introduce modalities: if we have the rule $a \Rightarrow_{\mathsf{O}} b$ and a holds, then we obtain $\mathsf{O}b$. That is to say, in the scenario where conditions described by a hold, the obligation of doing b is active as well. The advantage is that explicitly deriving modal literals such as

$\mathsf{O}b$ adds expressive power to the language, since $\mathsf{O}b$ may appear in the antecedent of other rules which, in turn, can be triggered.

3. Modal literals can only occur in the antecedent of rules. In other words, we do not admit nested modalities, i.e., rules such as $a \Rightarrow_{\mathsf{O}} \mathsf{P}b$. This is in line with our idea that the applicability of rules labelled with mode $\Box$ (where $\Box$ can be O for obligation, or P for permission) is the condition for deriving literals modalised with $\Box$.

4. The symbols O and P are not simple labels: they are modalities. O is non-reflexive (in a non-reflexive modal logic, $\Box a$ does not imply a, where $\Box$ is a modal operator): consequently, we do not have a conflict within the theory when $\neg a$ is the case and we derive that a is mandatory ($\mathsf{O}a$); this amounts to having a violation. The modality P works in such a way that two rules for P supporting a and $\neg a$ do not clash, but a rule like $\Rightarrow_{\mathsf{P}} b$ attacks a rule such as $\Rightarrow_{\mathsf{O}} \neg b$ and vice versa.

5. We should notice that permissive statements (e.g., $\mathsf{P}b$) can be derived by using permissive norms—these being usually called in the literature strong permissions—or, as done in Section 9, by showing that there no proof for the opposite obligation (i.e., given $\mathsf{P}b$, that there is no proof for $\mathsf{O}\neg b$)—these being the usual weak permissions.

6. The language thus includes two preference operators for obligations ($\otimes$) and permissions ($\odot$). As before,

$$Order \Rightarrow_{\mathsf{O}} PayBy7days \otimes Pay5\%Interest$$

means that, if you send an order, then paying your debts by 7 days is obligatory (i.e., $\mathsf{O}PayBy7days$ is obtained), but if you do not pay (i.e., $PayBy7days$ is factually the case), then you are obliged to pay with 5% interest.

For permissions, consider the U.S. Copyright Act, which states that the copyright owner may elect, at any time before final judgment is rendered, to recover, instead of actual damages and profits (the first option), an award of statutory damages for all infringements involved in the action (second option):

$$Owner \Rightarrow_{\mathsf{P}} ActualDamages \odot StatutoryDamages$$

If the US legal system allows for deriving in a specific case $\mathsf{O}\neg ActualDamages$ (using other norms), then we only conclude $\mathsf{P}StatutoryDamages$.

Ordered sequences of obligations or permissions can either be given explicitly, or inferred from other rules. However, we point out that, in domains such as the law, normative documents often explicitly contain provision with such structures.

Let us illustrate in some more details the scenarios just mentioned. Consider the Australian "National Consumer Credit Protection Act 2009" (Act No. 134 of 2009) which is structured in such a way that for every section establishing an obligation or a prohibition, the penalties for violating the provision are given in the section itself.

Example 10.1 (National Consumer Credit Protection Act 2009)**.** *Section 29 (Prohibition on engaging in credit activities without a licence) of the act recites:*

> *(1) A person must not engage in a credit activity if the person does not hold a licence authorising the person to engage in the credit activity.*
> *Civil penalty: 2,000 penalty units.*
> *[. . .]*
> *Criminal penalty: 200 penalty units, or 2 years imprisonment, or both.*

This norm can be represented as

$$r_1 : \Rightarrow_{\mathsf{O}} \neg \mathit{CreditActivity} \otimes \mathit{2000CivilPenaltyUnits}$$
$$r_2 : \mathit{CreditLicence} \Rightarrow_{\mathsf{P}} \mathit{CreditActivity}$$

where $r_2 > r_1$. The first rule states that, in absence of other information, a person is forbidden to engage in credit activities ($\mathsf{O}\neg \mathit{CreditActivity}$), and then the second rule establish an exception to the prohibition, or in other terms, it recites a condition under which such activities are permitted. The section continues by giving explicit exceptions (permissions) to the prohibition to engage in credit activity, even without a valid licence.

Remark 10.1. *This kind of structure has been successfully used for applications in the area of business process compliance [Governatori and Shek, 2012]. In a situation governed by the rule $\Rightarrow_{\mathsf{O}} a \otimes b$ and where $\neg a$ and b hold, the norm has been complied with (even if to a lower degree than if we had a). On the contrary, if we had two rules $\Rightarrow_{\mathsf{O}} a$ and $\neg a \Rightarrow_{\mathsf{O}} b$, then the first norm would have been violated, while the second would have been complied with. But the whole case would be not compliant*

[Governatori and Sadiq, 2008]. Consider the following example:

$$
\begin{array}{ll}
r_1: & \mathit{Invoice} \Rightarrow_{\mathsf{O}} \mathit{PayWithin7days} \\
r_2: & \mathsf{O}\mathit{PayWithin7days}, \neg \mathit{PayWithin7days} \Rightarrow_{\mathsf{O}} \mathit{Pay5\%Interest} \\
r_3: & \mathsf{O}\mathit{Pay5\%Interest}, \neg \mathit{Pay5\%Interest} \Rightarrow_{\mathsf{O}} \mathit{Pay10\%Interest}.
\end{array}
$$

What happens if a customer violates both the obligation to pay within 7 days after the invoice and the obligation to pay the 5% of interest, but she pays the total amount plus the 10% of interest? In the legal perspective the customer should be still compliant, but in this representation, contract clauses r_1 and r_2 have been violated. However, if we represent the whole scenario with the single rule

$$\mathit{Invoice} \Rightarrow_{\mathsf{O}} \mathit{PayBy7days} \otimes \mathit{Pay5\%Interest} \otimes \mathit{Pay10\%Interest},$$

then the rule is not violated, and the customer is compliant with the contract.

Section 29 shows that there are cases where the textual provisions of norms themselves suggest to represent the norms using sequences of obligations. However, there are other cases, witnessed by the invoice scenario above, where contrary-to-duties, e.g., sequences of obligations and their penalties, are stated in different norms. The framework proposed in this paper is agnostic about the format in which norms are modelled. It offers syntactic structures and reasoning mechanisms suitable to handle both cases. It is up to a legal knowledge engineer to decide which format is the most suitable for the needs of the application at hand. [Governatori and Rotolo, 2006] presents a sequent-style calculus to obtain rules with sequences of obligations (for example, to be used in a compliance application) from rules without them.

Sequences of permissions are a natural fit for expressions like "the subject is authorised, in order of preference, to do the following: (list)" or "the subject is entitled, in order of preference, to one of the following: (list)". This is illustrated in the next example, which offers a case of permissive right[36]:

[36]Hence, we speak here of *entitlements* or *rights*, as corresponding to options for exercising the same general permissive right to compensation. In this perspective, we can model them as permissions on one party (in this case the copyright owner) generating an obligation on another party (in this case the infringer). This is in line with the classic conception of rights proposed, for instance, in [Hohfeld, 1913; Hohfeld, 1917], which does not properly view them as powers: a power is typically required there to generate further normative effects (such as duties, juridical relations, etc.). For a more detailed discussion on these issues, see [Sartor, 2005].

Example 10.2 (U.S. Copyright Act)**.** *A concrete instance of sequences of permissions is given by Section 504(c)(1) (Remedies for infringement: Damages and profits) of the U.S. Copyright Act (17 USC §504).*

> *Except as provided by clause (2) of this subsection, the copyright owner may elect, at any time before final judgment is rendered, to recover, instead of actual damages and profits, an award of statutory damages for all infringements involved in the action, with respect to any one work, for which any one infringer is liable individually, or for which any two or more infringers are liable jointly and severally, in a sum of not less than $750 or more than $30,000 as the court considers just. [...]*

The above provision can be modelled as

$$Infringment, BeforeJudgment \Rightarrow_{\mathsf{P}} ActualDamages \odot StatutoryDamages$$

The above rendering of the textual provision is based on the interpretation of the term 'instead', which suggests that the copyright owners are entitled by default the award of the actual damages and profits, but they may elect to recover statutory damages, which is then the second option if exercised by the relevant party.

Here is another example showing a case where a sequence of permissions is used to derogate other obligations and prohibitions.

Example 10.3 (Italian Law 68/1999)**.** *Consider Article 4, Comma 4 of the Law 68/1999 (Right to work for people with disabilities).*

> *For [...] workers [who suffered minor disabilities as a result of injury or illness, when complying with their duties] the injury or illness does not allow for justified dismissal in case they can be employed for, and assigned to equivalent or job duties and tasks or, when this is not applicable, even to lower job tasks. In the case of allocation to lower tasks they have the right to retain the more favorable treatment corresponding to the original tasks. If any such employees cannot be assigned to equivalent or lower tasks, they are* ex officio *relocated, by the competent authorities [...], in other companies and assigned to activities compatible with their work capacity [...].*

This provision prohibits the termination of employment for workers suffering from a work related injury resulting in a minor disability. In case the employees are no longer able to carry out their normal job function

the comma permits to assign them to a different job position[37]*, with the option to terminate the employment if suitable job positions are not available. In addition it prescribes mechanisms to relocate workers whose employment has been terminated according to the previous condition.*

The comma can be formalised as follows:

$$r_1\colon \mathit{Injury}, \mathit{MinorDisability} \Rightarrow_{\mathsf{O}} \neg \mathit{TerminateEmployment}$$
$$r_2\colon \mathit{Injury}, \mathit{MinorDisability}, \neg \mathit{CurrentJob} \Rightarrow_{\mathsf{P}} \mathit{ChangeJob} \odot \mathit{TerminateEmployment}$$
$$r_3\colon \mathit{Injury}, \mathit{MinorDisability}, \mathit{TerminateEmployment} \Rightarrow_{\mathsf{O}} \mathit{Relocation}$$

where $r_2 \succ r_1$.

In case there is an available alternative position, the employer cannot terminate the employment, and has to assign the available position to the injured employee.

Suppose, now that the alternative (equivalent or lower) job position requires a licence, meaning that it is forbidden to perform the activities required without a licence:

$$r_4\colon \neg \mathit{License} \Rightarrow_{\mathsf{O}} \neg \mathit{ChangeJob}$$

with $r_4 \succ r_2$*. If the employee does not posses the required licence, then the employer can terminate the employment and has to refer the employee to the relevant authority for relocation.*

A knowledge base, a *Deontic Defeasible Theory* is as the previous section but it also includes a set of permissive rules:

$$\langle F, R^C, R^{\mathsf{O}}, R^{\mathsf{P}}, \prec \rangle \tag{17}$$

where R^C is a set of constitutive rules, R^{O} is a set of prescriptive rules, and R^{P} is a set of permissive rules. Of course, the set R^{O} can include rules with $\otimes$-expressions (such as $a \otimes b \otimes c$) in their consequent, while the set R^{P} can include rules with $\odot$-expressions (such as $d \odot \neg e$) in their consequent.

We do not provide full proof theory here, which anyway obeys the main intuition for the propositional basic deontic cases (see [Governatori *et al.*, 2013a]).

Informally, conditions for deriving obligations are the ones described in the previous section, except for the mechanism for $\otimes$ and the fact that we also have permissive rules. To show that l is defeasibly provable

[37] This in general might be prohibited based on contractual conditions of by other laws.

as an obligation, there are again two ways: (1) the obligation of l is a fact, or (2) l must be derived by the rules of the theory (either in the form $\Rightarrow_{\mathsf{O}} l$ or in the form $\Rightarrow_{\mathsf{O}} a_1 \otimes \cdots \otimes a_n \otimes l \otimes b_1 \otimes \cdots \otimes b_m$ where $\sim a_1 \ldots \sim a_n$ are provable using constitutive rules). In the second case, three conditions must hold: (2.1) l does not appear as not obligatory as a fact, and łq is neither provable as an obligation nor as a permission using the set of modal facts at hand; (2.2) there must be a rule introducing the obligation for l which can apply; (2.3) every rule s for $\sim l$ is either discarded or defeated by a stronger rule for l. If s is an obligation rule, then it can be counterattacked by any type of rule; if s is a defeater or a permissive rule, then only an obligation rule can counterattack it.

As for permissions, l can be derived as permitted by $-\partial_{\mathsf{O}} \sim l$ or from the rules of the theory (either in the form $\Rightarrow_{\mathsf{P}} l$ or in the form $\Rightarrow_{\mathsf{P}} a_1 \odot \cdots \odot a_n \odot l \odot b_1 \odot \cdots \odot b_m$ where $\mathsf{O}\sim a_1 \ldots \mathsf{O}\sim a_n$ are provable). Hence, proof conditions differ from their counterpart for obligation in two aspects: scenarios where both $+\partial_{\mathsf{P}} l$ and $+\partial_{\mathsf{P}} \sim l$ can hold, but $+\partial_{\mathsf{O}} \sim l$ must not hold; any applicable rule s attacking l as permitted and supporting $\sim l$ can be counterattacked by any type of rule t supporting l, since s must be an obligation rule, and permissive rules can only be attacked by obligation rules.

We intuitively illustrate how logic works with the help of a simple abstract example.

Example 10.4. *Consider the following theory D:*

$$\begin{aligned} F &= \{a, e, \mathsf{O}z\} \\ R^C &= \{r_1 : a \Rightarrow b, \\ &\qquad r_2 : b \Rightarrow \neg p\} \\ R^{\mathsf{O}} &= \{r_3 : e \Rightarrow_{\mathsf{O}} p \otimes \neg q \\ &\qquad r_4 : a, \mathsf{O}z \Rightarrow_{\mathsf{O}} s\} \\ R^{\mathsf{P}} &= \{r_5 : e, b, \mathsf{P}s \Rightarrow_{\mathsf{P}} q \odot t\} \\ \prec &= \{\langle r_3, r_5\rangle\}. \end{aligned}$$

Let us start with the facts. If we consider rules in R^C, only fact a plays a role. No conflict occurs here, so we can derive (in addition to $+\partial_C a$ which is trivially obtained from a being a fact) $+\partial_C b$, and $+\partial_C \neg p$.

Let us move to the deontic rules. Fact e triggers r_3, for which there is no conflict, so one can derive $+\partial_{\mathsf{O}} p$ (i.e., $\mathsf{O}p$). However, we also derived $+\partial_C \neg p$, which amounts to the violation of the primary obligation, so we could obtain $+\partial_{\mathsf{O}} \neg q$. Can we? Indeed, r_3 defeats r_5 (since it is stronger according to $\prec$), which in turn is applicable because $+\partial_C a$ and $+\partial_C b$ are the case, and because $-\partial_{\mathsf{O}} \neg s$ is derivable from the fact that r_4 is applicable and is not attacked by any other rules; so $+\partial_{\mathsf{O}} s$ is the case as well as $-\partial_{\mathsf{O}} \neg s$ and so $+\partial_{\mathsf{P}} s$. Hence, from r_5 we can only derive $+\partial_{\mathsf{P}} t$.

Finally, let us recall some results of the system, which are conceptually relevant for a deontic point of view.

Theorem 10.1. *Let D be an O-consistent Defeasible Theory, i.e. a theory such $\prec$ is acyclic and for any literal l, the set of facts F does not contain any of the following pairs: $\mathsf{O}l$ and $\mathsf{O}{\sim}l$, $\mathsf{O}l$ and $\mathsf{P}{\sim}l$.*

For any literal l, it is not possible to have both $D \vdash +\partial_{\mathsf{O}} l$ and $D \vdash +\partial_{\mathsf{O}} {\sim}l$.

Theorem 10.2. *Let D be an O-consistent Defeasible Theory. For any literal l:*

1. *if $D \vdash +\partial_{\mathsf{O}} l$, then $D \vdash -\partial_{\mathsf{O}} {\sim}l$;*
2. *if $D \vdash +\partial_{\mathsf{O}} l$, then $D \vdash -\partial_{\mathsf{P}} {\sim}l$;*
3. *if $D \vdash +\partial_{\mathsf{P}} l$, then $D \vdash -\partial_{\mathsf{O}} {\sim}l$.*

Definition 10.1. *Let D be a Defeasible Theory. A literal l is* weakly permitted *iff $D \vdash -\partial_{\mathsf{O}} {\sim}l$.*

Corollary 10.1. *Let D be any O-consistent Defeasible Theory. For any literal l, if $D \vdash +\partial_{\mathsf{O}} l$, then l is weakly permitted.*

10.3 Further readings

For an in-depth presentation of Defeasible Deontic Logic, its properties and a detailed analysis of how to use it to model obligations and permissions (and several ways to do it) we refer the reader to [Governatori *et al.*, 2013a]. Notice that the full logic still preserves the computational complexity of the propositional case: the set of conclusions of any theory D can be computed in time linear to the size of D.

In Section 5.2 we mentioned two fundamental roles of positive permissions acknowledged in the literature: stating exceptions to obligations, and preventing a legislator from issuing future obligations. In [Governatori *et al.*, 2013a] the second case is not explicitly discussed, but it can be easily handled by introducing strict rules for permission, or by assuming that certain defeasible rules for permission are stronger than any other rule. An explicit treatment, in a different logic, is offered in [Makinson and van der Torre, 2003].

We would like to recall another approach to the representation of deontic notion in defeasible contexts, originally proposed by [Prakken, 1996]. This approach combines a suitable rule-based nonmonotonic logic (in the above paper default logic) with a modal deontic logic by choosing that modal as the underlying logical language and by encoding the semantics of the modalities in the strict rules. This allows for a natural distinction between strong and weak permission: p is strongly permitted if $\mathsf{P}p$ is (nonmonotonically) derivable and p is weakly permitted if $\mathsf{O}{\sim}p$ is not nonmonotonically derivable.

11 Institutionalised agency and normative positions

In Section 4, a number of legal notions have been mentioned and we also recalled the well-known theory of normative positions. An influential part of the literature (reported in that section) has argued that several types of normative positions (such as the limited, but quite investigated Hohfeldian sets) can be modelled by combining suitable deontic logics, logics of action, and a logic for constitutive rules (to model the idea of power).

Governatori and Rotolo [2008b] collected all those intuitions into a single modal extension of Defeasible Logic. This section presents elements of this framework.

11.1 Introduction

The background of extension of Defeasible Logic is Kanger-Lindahl-Pörn [Kanger, 1957; Lindahl, 1977; Pörn, 1977] theoretical account of organised interaction (see [Elgesem, 1997]). The basic idea is to describe agents' interaction within a multi-modal logical setting. The resulting view is abstract but flexible, as social agency is captured in a modular way by simply combining different modal operators.

Despite some limitations, modal logic of agency [Elgesem, 1997] is a useful tool[38] thanks to its flexibility, as actions are simply taken to be relationships between agents and states of affairs. We focus on two well-known agency notions. The first is the idea of personal and direct action to realise a state of affairs, which is formalised the modal operator E, such that a formula like $E_i p$ means that the agent i brings it about that p. Different axiomatisations have been provided for it [Governatori and Rotolo, 2005]. Here we consider two basic logical properties of this operator[39]:

$$E_i A \to A \qquad (\mathbf{T})$$

$$E_i E_j A \to \neg E_i A \qquad (\mathbf{EE\neg E})$$

Schema **T** expresses the successfulness of actions that is behind the common reading of the "bring about" concept. Schema (**EE¬E**) is a specific axiom advanced, for example, in [Santos *et al.*, 1997]. The brings-it-about operator expresses actions performed directly and personally. Hence, (**EE¬E**) states a principle of rationality for modelling co-ordination in institutional organisations: it is counter-intuitive that the same agent brings it about that p and brings it about that somebody else achieves p.

The second aspect of agency is that of attempt, formalised by the operator H [Santos *et al.*, 1997]. $H_i p$ says that i attempts to make it the case that p. The operator H_i is not necessarily successful. Here we simply assume that each successful action is also an attempt (see [Santos *et al.*, 1997]):

$$E_i p \to H_i p \qquad (18)$$

Let us focus now on the idea of institutionalised power. As we said in Section 4, this notion is central in legal theory and comes from the distinction between the practical ability to realise a state of affairs

[38] See Chapter 5, Volume I, of this Handbook.

[39] Besides these schemata, the logic for E is usually closed under logical equivalence. Other common properties, which are not considered here, correspond to $\neg E_i \top$ (No) and $(E_i A \wedge E_i B) \to E_i(A \wedge B)$ (C).

[Elgesem, 1997; Governatori and Rotolo, 2005] and the institutional power to do this [Makinson, 1986]. For example, if in an auction i raises one hand, this implies that the act of making a bid is also obtained. In principle, this kind of ability should be distinguished from the practical capacity to obtain a certain state of affairs. The attempt to make a bid may not be successful: its being successful, within the institutional context (the auction), depends on whether that institution makes it effective. It is up to institutional (constitutive) *rules* to establish whether i's act makes so that a bid is effective or not, namely, that i's act *counts as* bidding.

The logical nature of this kind of rules has been investigated following different directions starting from the seminal work by [Jones and Sergot, 1996] (see Chapter 6, Volume I, of this Handbook and also [Gelati *et al.*, 2004]). Many of these approaches explicitly recognise that constitutive rules are defeasible. In fact, it is intuitive that, e.g., if the agent i raises one hand, this may count as making a bid but this does not hold if i raises one hand *and* scratches his own head.

As we have argued in Section 9 (see [Governatori and Rotolo, 2008b]), a simple way is to model constitutive rules is through propositional Defeasible Logic.

Notice that the framework we have just recalled is able to capture some composite normative concepts. In particular, [Gelati *et al.*, 2004] show that the introduction of the notion of proclamation allows to account for the ideas of delegation. The logical representation of these ideas has a counts-as structure. Institutional proclamations are formalised by the modal operator *proc*: the expression $proc_i p$ means that agent i proclaims p[40] The combination of *proc*, agency operators, and constitutive rules enables us to capture two forms of normative delegation, intended as kinds of true representation [Gelati *et al.*, 2004]. The first is $proc_j(proc_i A) \Rightarrow_C E_j(proc_i A)$, that is, when j proclaims that i proclaims that p, this counts as j's making so that i proclaims that p[41]. In addition, we can have $proc_j(E_i p) \Rightarrow_C E_j(E_i p)$. This type of representation is necessary when the representative substitutes a principal which would not be able to perform directly the activity which is delegated to the representative.

[40] As is well-known, agent communication concepts play an important role in modelling agent coordination. In [Gelati *et al.*, 2004] the speech act of proclaiming has been defined to capture some minimal properties of all speech acts that are intended to modify the institutional world.

[41] Of course, the achievement of p will depend on the presence on another rule which states that $proc_i p$ counts as $E_i p$.

11.2 The framework and possible developments

Since we want to reason about actions we extend the language of Defeasible Logic with a set of action symbols; we will use $\alpha_i, \beta_i, \gamma_i$ to denote atomic actions. The meaning of an action symbol, for example α_i, is that the action corresponding to it has been performed by agent i, while we use $\neg\alpha_i$ to denote that the action described by α_i has not been performed. Given the modal operators E_i, H_i, and $proc_i$ we form new literals as follows: i) if l is a literal then $proc_i l$ is a literal; ii) if l is a literal then $E_i l$, $\neg E_i l$, $H_i l$ and $\neg H_i l$ are literals if l is different from $E_i m$, $\neg E_i m$, $H_i m$ and $\neg H_i m$, for some literal m.

If we confine the framework to institutional agency only, a Defeasible Institutional Action Theory is a structure

$$I = (A, F, R^C, \{R^i\}_{i\in A}, \prec)$$

where, A is a finite set of agents, F is a set of facts, R^c is a set of constitutive rules, $\{R^i\}_{i\in A}$ is a family of sets of results-in rules (i.e., $\rightarrow^i$, $\Rightarrow^i$, $\rightsquigarrow^i$, $\forall i \in A$), and $\prec$, the superiority relation, is a binary relation over the set of rules (i.e., $\prec \subseteq (R^c \cup R^A)^2$), where $R^A = \bigcup_{i\in A} R^i$.

The intuition is that, given an institution, F consists of the description of the institutional facts[42], either in form of states of affairs (literal and modal literal) or actions that have been performed. R^C describes the basic inference mechanism internal to an institution, while R^A encodes the transitions from state to state occurring as the results of actions performed by the agents within the organisation. As previously done with deontic operators, the rules in R^A are used to introduce modal operators. To capture these notions we impose some restrictions on the form of rules: literals of the form $E_i l$, $\neg E_i l$, $H_i l$ and $\neg H_i l$ are not permitted in the consequent of results-in rules for i, while actions symbols are not permitted in the consequent of results-in rules. The first restriction is motivated from the fact that 1) results-in rules are the rules to introduce the modalities and in the present context sequences of modalities for the *same* agent are useless[43] 2) constitutive rules make possible the derivation of institutional actions (modalised literals) only when they follow from specific actions (intentionally) performed by the agent. The second restriction is due to the idea that results-in rules describe, as their name suggests, the results of actions, not actions themselves.

Let us see by means of some examples the intuition behind this formalism. Suppose the agent i is acting in the context of an auction.

[42] For the notion of institutional fact, see Section 4.2.

[43] An expression like $E_i E_i A$ is useless since it is equivalent to $E_i A$.

Then we may have cases like the following[44]:

$$\mathbf{bids_i},\ \mathit{auction\_begun} \Rightarrow^i \mathit{offer} \tag{19}$$

This rule is an example corresponding to the introduction of the modality E_i. In fact, agent i's fulfilment of the conditions in the antecedent produces the occurrence of *offer*: agent i's action of bidding has the result that i has made an offer. If *offer* can be derived, this permits the introduction of $E_i(\mathit{offer})$.

$$\mathit{auction\_begun} \Rightarrow^i \neg \mathit{offer} \tag{20}$$

The example above does not specify any action in the antecedent (empty action). This means that, when the auction is begun, agent i's refraining from doing any action (generic omission) has the result to have no offer. In logical terms, also this case can lead to the introduction of E[45].

Now suppose that agent i is acting on behalf of agent j.

$$\mathbf{bids_i},\ \mathit{proc}_i(E_j\,\mathit{offer}) \Rightarrow^j \mathit{offer} \tag{21}$$

This formula means that the fact that agent i makes a bid and proclaims that agent j makes the offer permits to introduce E_j, namely that $E_j\,\mathit{offer}$.

Let us consider examples of constitutive rules.

$$\mathbf{raises\_hand_i},\ \mathit{auction\_begun} \Rightarrow_C \mathbf{bids_i} \tag{22}$$

This rule says that that agent i's action of raising one hand counts as agent i's action of bidding, when the auction is begun.

$$\mathit{auction\_begun},\ E_i(\mathit{offer}) \Rightarrow_C \neg\mathbf{raises\_offer_i} \tag{23}$$

Also here we have agent i's generic refraining from doing any action in the antecedent. This example represents the institutional connection linking such refraining, and *the fact* that agent i made an offer when the auction is begun, to agent i's specific refraining from raising a new offer. Notice that the same meaning is assigned to counts-as rules where the antecedent contains only non-modal literals.

$$\mathit{auction\_begun},\ \mathbf{raises\_hand_i} \Rightarrow_C \mathit{offer} \tag{24}$$

[44]Bold type expressions correspond to action symbols, the italicised ones to state of affairs.

[45]The ideas of empty action and refraining from doing a specific action should not be confused with what it is expressed by $\neg E_i A$. As we will see, this last corresponds to the non-derivability of A within I, which can depend also on reasons that have nothing to do with agent i's refraining from acting to realise A.

This rule is an example of the institutional analogous of results-in rules, where an action and a state of affairs occur respectively in their antecedent and consequent. However, in this case the result is an institutional fact and follows by convention only within the institution. In fact, that an offer is a consequence of agent i's raising one hand is not a simple matter of agent i's action results. The attempt of agent i to make an offer by raising the hand is effective only if the institution recognises this.

Let us see a couple of examples with more than one agent. As above, agent i is acting on behalf of agent j.

$$proc_i(E_j\, offer) \Rightarrow_C E_i(E_j\, offer) \tag{25}$$

This rule says that if agent i proclaims that agent j makes an offer, then this counts as agent i brings it about that agent j makes such an offer.

$$proc_i(E_j\, offer), \mathbf{raises\_hand}_i \Rightarrow_C \mathbf{bids}_j \tag{26}$$

Rule (26) expresses that agent i's proclamation that agent j makes an offer counts as agent j's action of bidding.

The logic developed in [Governatori and Rotolo, 2008b] did not include deontic operators. Extending it with deontic components (such as those in the previous section) is easy, since Modal Defeasible Logic is modular. In particular, following [Sartor, 2005] and in order to correctly capture significant types of basic normative positions (such as those mentioned in Section 4.4) we have to introduce directed obligations such as O^i: if i is an agent, an expression like $\mathsf{O}^i p$ means that p is obligatory and i is the beneficiary of this obligation. If combined with the formalism of E (of the "bring about" concept), we can write rules like

$$E_{Tom} Purchase\_Car \Rightarrow_{\mathsf{O}}^{Mary} E_{Tom} 12,000 EUR$$

which means "if Tom buys a car from Mary, then it is obligatory, towards Mary, that Tom pays 12,000 euros to her". In other words, we write prescriptive rules as in the Defeasible Deontic Logic presented in the previous sections, but we add as the superscript of the arrow the beneficiary of the obligation we obtain through the rule. Hence, if applicable, this rule allows for deriving $\mathsf{O}_{\mathsf{O}}^{Mary} E_{Tom}$. This is a case of *obligative right* (in [Sartor, 2005]'s sense), i.e., a right in Hohfeld's terminology.

Following the seminal intuition of [Jones and Sergot, 1996] and the further developments by [Gelati *et al.*, 2004], *potestative judgements* can be naturally captured by using constitutive rules. A rule like

$$proc_{President} State\_Emergency \Rightarrow_C E_{President} State\_Emergency$$

means "the president has the power to declare a state of emergency", i.e., expresses a form of *declarative power*.

Accordingly, the definition of a Defeasible Institutional Action Theory can be extended as follows:

$$I = (A, F, R^C, \{R^i\}_{i \in A}, \{R^i_{\mathsf{O}}\}_{i \in A}, \prec)$$

where $\{R^i_{\mathsf{O}}\}_{i \in A}$ is a family of sets of rules for directed obligations.

Proofs conditions are omitted [**?**, for which see]]AILaw08. Here we only explain how conflicts between rules can be detected and illustrate the machinery with a simple example.

Rules collide if they support the conclusion of complementary literals. In standard Defeasible Logic two literals are complementary to each other if one is the negation of the other. This means that the two literals cannot hold at the same time. The extension with modal operators as those discussed above has to consider when modal literals are in conflict with each other. In particular, since the agency operator E is successful (i.e., $E_i l \rightarrow l$), it is not possible to have together $E_i l$ for some agent i and $\sim l$. In a similar way we have to capture the strong notion of agency we intend to model within our framework, i.e., where $E_i E_j l \rightarrow \neg E_i l$.

Given an atomic literal p, Ep denotes any string $E_{i_1} \ldots E_{i_n} p$ where $E_{i_1} \ldots E_{i_n}$ is a (possibly empty) string of positive modal operators such that $\forall 1 \leq j < n, i_j \neq i_{j+1}$. Let l be a literal, $\mathcal{C}(l)$ denotes the complement of l, i.e., the set of literal that cannot be true when l is.

- if $l = p$, then $\mathcal{C}(l) = \{E{\sim}p\}$;
- if $l = E_i p$, then $\mathcal{C}(l) = \{E{\sim}p, E{\neg}E_i p\}$;
- if $l = \neg E_i p$, then $\mathcal{C}(l) = \{E E_i p\}$.

The meaning of the first condition is that if p is true then no agent prevented p; for the second condition we have that if an agent i has realised p, then no other agent prevented p and no agent prevented i from realising p. Finally if an agent i has refrained from doing p, then it is not possible that some other agents achieved that i did p.

The definition of complementary literals is trivial. We just notice, as expected, that $\mathsf{O}^i p$ and $\mathsf{O}^j \neg p$ are compatible.

Also, the proof theory embeds the notion of rule conversion. In this context, this means that a constitutive rule can be used as it were a results-in rule if all the literals occurring in its antecedent are proved as appropriate results-in conclusions. We thus say that we have a *conversion* from a constitutive rule into a results-in rule. For example, suppose we have that

$$\textit{auction\_begun},\ \mathbf{raises\_hand_i} \Rightarrow_C \textit{offer}$$

If we have **raises_hand$_i$** and prove *auction_begun* as a results-in conclusion, in particular as E_i*auction_begun*, then we can say that agent i brings *offer* about, namely that E_i*offer* (see [Governatori and Rotolo, 2008a]).

Let us now consider how to represent the following business scenario. For normal orders a company has pre-defined invoices and the finance department can delegate the preparation of the invoices to the shipping department. The preparation of an invoice requires to check that the details in it are correct and to sign it. However special orders require more care and processing, and the finance department is in charge for their invoices. Goods can be delivered only after the finance department has prepared the invoice. Finally, when the buyer pays the invoice, the shipping department is obliged to deliver the goods in favour of the buyer. This scenario is depicted by the following institutional theory,

$r_1 : proc_F(E_S(invoice\_ready)), E_S(invoice\_ready) \Rightarrow^F invoice\_ready$
$r_2 : special\_order, E_S(invoice\_ready) \Rightarrow_C \neg invoice\_ready$
$r_3 : \mathbf{sign\_invoice}_X \Rightarrow^X invoice\_checked$
$r_4 : invoice\_checked \Rightarrow_C invoice\_ready$
$r_5 : E_F(invoice\_ready) \Rightarrow_C ship\_order$
$r_6 : E_B(paid\_invoice) \Rightarrow^B_{\mathsf{O}} E_S deliver$

where $r_1 \prec r_2$ and $r_4 \prec r_2$. Here rule r_1 is the rule governing the delegation of the preparation of the invoice, where r_2 is an exception to it. r_3 is a schema that establishes that the act of signing an invoice by an agent (a role) X results in the invoice being checked by X. The meaning of r_4 is that according to the business rule of the company is that once an invoice has been checked then the invoice is ready to be sent. Finally r_5 states that items can be shipped only after their invoice has been approved by the finance department.

Let us consider the following scenario. The company receives an order. The finance department considers the order to be a standard order and it delegates the whole process to the shipping department, which processes it and a clerk in this department signs the invoice. In this case the facts are $proc_F(E_S(invoice\_ready))$, and $\mathbf{sign\_invoice}_S$. We can apply r_3 to derive $E_S(invoice\_checked)$. According to rule r_4 we have that the invoice is ready. However the invoice has been signed by a clerk in the shipping office, the result of this action is qualified as an act performed by the shipping department. This means that we carry over the qualification from the antecedent to the consequent of rule r_4. Hence we obtain $E_S(invoice\_ready)$. Since the shipping department was delegated by the finance department to process the invoice, we can

apply rule r_1 to derive that the invoice had been prepared by the finance department via delegation ($E_F(invoice\_ready)$) and the order can be delivered.

On the other hand, if an order is classified as a special order, then the only alternative is that the finance department process the invoice by itself, that is somebody in the finance department has to sign the invoice.

Finally, if $E_B(paid\_invoice)$ we can derive $\mathsf{O}^B E_S deliver$.

11.3 Further readings

The literature background of the framework presented above, regarding the combination of deontic and action logics, is offered in Section 11.1. Of course, the literature on this topic is immense and it is out of the scope of this chapter to discuss it: see Chapter 5, Volume I, of this Handbook for a comprehensive overview.

For an in-depth presentation of Defeasible Institutional Agency and its properties we refer the reader to [Governatori and Rotolo, 2008b]. This still preserves the computational complexity of the propositional case: the set of conclusions of any theory D can be computed in time linear to the size of D. In section above we briefly mentioned the idea of *rule conversion*. In general, this idea allows for modelling peculiar interactions between consequence relations or different modal operators. In general, notice that in many formalisms it is possible to convert from one type of conclusion into a different one. Take for example the right weakening rule of non-monotonic consequence relations

$$\frac{B \vdash C \quad A \mathrel{|\!\sim} B}{A \mathrel{|\!\sim} C}$$

which allows for the combination of non-monotonic and classical consequences. Indeed, one may consider that each rule type in any Modal Defeasible Logic corresponds to a specific consequence relation. A comprehensive discussion of the concept of conversion can be found in [Governatori and Rotolo, 2008a].

12 Legal provisions and legal norms: Interpretation and deontics

12.1 Introduction

Legal doctrine and judicial practice distinguish among a number of canons for interpreting legal statutes, i.e., different rules that are employed in legal systems as patterns for constructing arguments aimed at justifying certain interpretations, while attacking other interpretations. MacCormick and Summers [1991], summarising the outcomes of a vast study on statutory interpretation, involving scholars from many different legal systems, distinguish eleven types of arguments. A different list of interpretive arguments was developed by [Tarello, 1980] and identifies fourteen types of arguments. Here are some examples:

Argument from ordinary (or literal) meaning: a statutory provision should be interpreted according to the (literal) meaning a native speaker of a given language would ascribe to it.

Argument from contextual harmonisation: a statutory provision should be interpreted in light of the whole statute it is part of, or in light of other statutes it is related to.

Argument from precedent: a statutory provision should be interpreted in conformity with previous interpretations.

Argument from analogy: if a statutory provision is similar to provisions of other statutes, then it should be interpreted to preserve the similarity of meaning.

Argument from substantive reasons: if a fundamental goal can be promoted by one rather than another interpretation of a statutory provision, then the provision should be interpreted in accord with the goal.

Argument from intention: if a legislative intention concerning a statutory provision can be identified, the provision should be interpreted in line with that intention.

Following [Prakken and Sartor, 2013; Macagno *et al.*, 2012], in this section of Part III we recall an extension of Defeasible Logic which was devised by [Rotolo *et al.*, 2015] for modelling reasoning about interpretive canons and thus for justifying the choice of a certain canon and the resulting legal outcome over competing interpretations. [Sartor *et al.*,

2014] argued that an interpretive canon for statutory law can be expressed as follows: if provision n occurs in document D, n has a setting of S, and n would fit this setting of S by having interpretation a, then, n ought to be interpreted as a. For instance, the ordinary language canon has the following structure:[46] if provision n, stating that "Killing a man is punishable by no less than 21 years in prison", occurs in document D = Penal code, n has a setting of ordinary language, and n would fit this setting of ordinary language by having interpretation a = "Killing an adult male person is punishable by no less than 21 years in prison", then n ought to be interpreted as a.

The use of Defeasible Logic for modelling legal interpretation is based on following intuitions.

Intuition 1 (Reasoning and canons)**.** *We analyse the logical structure of interpretive arguments (in the sense of [MacCormick and Summers, 1991]) using a rule-based logical system. In particular, interpretation canons are represented by defeasible rules, where*

- *antecedent conditions of interpretation rules can be of any type (assertions, obligations, etc.), including the fact that another canon is refuted or that another legal provision ought to be interpreted in a certain way;*
- *the conclusion of interpretation rules is an interpretive act leading to an interpretation of a certain provision n and thus to a sentence which expresses the result of such an interpretation and paraphrases n [Brozek, 2013]. If n and n' are legal provisions, the following is an example of interpretation rule regarding n':*

IF
n ought to be interpreted literally as a, AND
n is related with n', AND
a entails a',
THEN
n' is interpreted by coherence as a'.

We use these rules to devise a reasoning machinery that mirrors legal reasoning about interpretive canons. The resulting rule-based system is in line with the basic ideas inspiring the argumentation system by [Prakken and Sartor, 2013].

Notice that the above intuition distinguishes the interpretive act from the result of the interpretation:

[46] This argument supports the option that a provision be interpreted according to the meaning a native speaker of a given language would ascribe to it.

Intuition 2 (A- and O-interpretation)**.** *We assume the distinction between interpretation as activity and as outcome [Ross, 1958, p. 117] (see [Tarello, 1980, p. 39]):*

- *interpretation as activity (*A-interpretation*) (literal or from ordinary language, by coherence, etc.) views any argumentative canon as a means through which a certain meaning is ascribed to a legal provision, and*
- *interpretation as outcome (*O-interpretation*) is precisely the meaning obtained through a certain interpretive act and ascribed to the provision.*

The distinction between interpretation as activity and as outcome is well known in continental legal theory, and it was introduced precisely to capture cases where, e.g., one has legal reasons to prefer a certain interpretive canon over others even though all considered canons support the same interpretive outcome. In other words, an interpretive act I *of n as a (A-interpretation of n) is a way to bring about that a (O-interpretation of n) is the case. For example, in Intuition 1, the A-interpretation of n' is the act interpreting n' by coherence, while the resulting O-interpretation from that act is a', i.e., a sentence expressing the meaning attributed to n' through the interpretation by coherence.*

Accordingly, this approach takes stock of the idea mentioned in Definition 4.1, Part II of this chapter.

Since different competing canons can be employed, different conflicting rules can be accordingly applied for interpreting statutes. Interpretation rules are thus defeasible. As argued in [Sartor *et al.*, 2014], some priority criteria should be applied to interpretation rules [Alexy and Dreier, 1991]. Such criteria impose preference relations over conflicting interpretive acts and outcomes. In other words, to address interpretive conflicts, we need to assume that one of the conflicting arguments is stronger than its competitors. Some legal traditions provide indeed general criteria for addressing conflicts of arguments on the basis of their priorities: for instance, several continental legal systems explicitly state that literal interpretation ought to be preferred, or that an argument concerning constitutional values ought to prevail over a historical argument (e.g., an argument based on the intent of the historical legislator).

However, ranking among interpretive acts and canons can be applied also when such acts are not in conflict. Suppose, for example, that provision n can be interpreted as a by adopting an argument by analogy and one from substantive reasons (see above); if n is a provision of criminal law (but analogy is admissible whenever it favours the defendant), then the argument from substantive reasons ought to be preferred, even though

both lead to read n as a.

Intuition 3 (Preferences over interpretations)**.** *A standard superiority relation over interpretation rules can be introduced to handle and solve conflicts between different interpretation rules. Consider the following example:*

Rule1
IF
n *ought to be interpreted literally as* a, AND
n *is related with* n', AND
a *entails* a',
THEN
n' *is interpreted by coherence as* a'

Rule2
IF
n'' *ought to be interpreted literally as* $\neg a$, AND
n *is related with* n'', AND
$\neg a$ *entails* $\neg a'$,
THEN
n' *is interpreted by coherence as* $\neg a'$.

Here, we can handle the conflict by stating, e.g., that **Rule1** $\prec$ **Rule2** *(or vice versa).*

Ranking among interpretive acts can be applied also when such acts are not in conflict. In this perspective, [Rotolo et al.*, 2015] made a different use—and proposed a different interpretation—of the operator* $\otimes$ *illustrated in the previous section: indeed,* $\otimes$ *can be employed to make explicit in single rules this idea. For instance,*

IF
n *ought to be interpreted literally as* a, AND
n *is related with* n'
THEN
n' *is interpreted by coherence as* $a' \otimes$
$\otimes$ n' *is interpreted by analogy as* a'

means that the most preferred interpretation resulting in a *is the one by coherence, but, if this is refuted, the second option is the interpretation by analogy. This does not require to only derive one interpretation resulting in* a *(other rules could first support interpretation by analogy of* n*) (for a technical study, see [Calardo* et al.*, 2018]).*

Following some doctrinal and judicial practice, [Sartor *et al.*, 2014] argued that interpretive canons are defeasible rules licensing deontic interpretive claims, namely, the claim that a certain expression in a statute ought, ought not, may or may not be interpreted in a certain way. For example, art. 12 in the general provisions of the Italian civil code states that the literal interpretation of statues ought to be preferred and this option is nothing but an interpretive prescription. Here, we follow this intuition with some adjustments.

Intuition 4 (Obligatory and admissible (permitted) interpretations)**.** *An interpretation can be* admissible *or* obligatory*. In the case of A-interpretations, for instance, an interpretive act* I *of* n *(A-interpretation of* n*) is admissible, if it is provable using a defeasible interpretation rule; it is obligatory, if this interpretation of* n *is the only one admissible. Similarly for O-interpretations. Indeed, consider the general provisions of the Italian civil code, which state at art. 12 that literal interpretation* I_{lit} *ought to be preferred: this would support that such interpretation is obligatory, unless another interpretation prevails. We have two options here:*

- *other conflicting interpretations can be derived, thus requiring to check if literal interpretation overrides the other options; if it does not, then the interpretation at stake is not even admissible;*
- *other non-conflicting interpretations can be provable; if they are, the interpretation at stake is only admissible, otherwise, it is obligatory.*

In other words, some basic form of deontic reasoning is relevant also for legal interpretation as such.

On the basis of the above intuitions, we can offer two options for modelling reasoning about interpretations: a defeasible logic for reasoning about the interpretation of abstract, non-analysed provisions and of structured provisions.

12.2 Deontic defeasible logics for legal interpretation

This framework handles the overall meaning of legal provisions intended as argumentative, abstract (i.e., non-analysed) logical units. In other words, a provision n is taken in its sentential entirety for interpretive purposes, i.e., as a non-analysed sentence without considering its internal (logical) structure. The following basic components (among others) are introduced:

- a set NORM $= \{n_1, n_2, \dots\}$ of legal provisions to be interpreted;

- as set $Prop = \{a, b, \dots\}$ of literals, corresponding to any sentences, which can be used to offer a sentential meaning to any provision n (a literal a is the meaning of provision n);
- a set $\mathsf{INTR} = \{\mathcal{I}_1, \mathcal{I}_2, \dots\}$ of interpretative acts or interpretations (literal interpretation, teleological interpretation, etc.) that return for any legal provision a sentential meaning for it;
- the deontic operators O, P, where O is the modality for denoting obligatory interpretations and interpretation outcomes and P for denoting the admissible ones;
- a set of rules like $a_1, \dots, a_n \Rightarrow^I C$ encoding interpretive arguments (i.e., rules that state what interpretive act can be obtained under suitable conditions); these rules expresses modes of reasoning within any given legal system;
- [Rotolo *et al.*, 2015] also introduce another type of rules, the so-called meaning rules encoding the ontology of legal and ordinary concepts. In other words, if a rule of this type states that if *Car* then *Vehicle*, this means that any car is a vehicle. Following [Boella *et al.*, 2010], we use constitutive rules to do this job.

Any interpretative act $\mathcal{I}_j$ in INTR can be thought as a function mapping one provision into one meaning. To keep notation compact and simple, we use $\mathcal{I}_j(n, a)$, instead of $\mathcal{I}_j(n) = a$, to say that $\mathcal{I}_j$ assigns a as meaning of n.

Let us see an example to illustrate how rules are.

Example 12.1. *Consider the following provision from the Italian penal code:*

> *Art. 575. Homicide. Whoever causes the death of a* man *[*uomo*] is punishable by no less than 21 years in prison.*

Consider now that paragraph 1 of art. 3 of the Italian constitution reads as follows:

> *Art. 3. All citizens have equal social status and are equal before the law, without regard to their sex, race, language, religion, political opinions, and personal or social conditions.*

The interpretation I_s *(interpretation from substantive reasons*[47]*) of art. 3 leads to c, which corresponds to the following sentence:*

[47] An argument from substantive reasons states that, if there is some goal that can be considered to be fundamentally important to the legal system, and if the goal can be promoted by one rather than another interpretation of the statutory provision, then the provision should be interpreted in accord with the goal.

> *All* persons *have equal social status and are equal before the law, without regard to their sex, race, language, religion, political opinions, and personal or social conditions.*

The following interpretation defeasible rule could be:

$$r_1: \quad kill\_adult, kill\_female, \mathsf{O}\,\mathsf{I}_s(art.3, c) \Rightarrow^I \mathsf{I}_c(art.575, b)$$

where b "Whoever causes the death of a person *is punishable by no less than 21 years in prison". In other words, if art. 3 of the Italian constitution states formal equality before the law without regard also to gender identity, then b is the best interpretation outcome of art. 575 of the penal code, with* I_c *denoting, for example, interpretation by coherence.*

Since interpretation rules can collide—i.e., interpretative arguments can be incompatible—it is fundamental to establish when interpretative acts are in conflict. More specifically, the *complementary* $\sim\phi$ *of an interpretation* ϕ is defined as follows:

ϕ	$\sim\phi$
$\mathsf{I}_i(n, a)$	$\sim\mathsf{I}_i(n, a) \in \{\neg\mathsf{I}_i(n, a), \mathsf{I}_i(n, b), \mathsf{I}_j(n, c) \mid a \neq b, a \neq c\}$
$\neg\mathsf{I}_i(n, a)$	$\sim\neg\mathsf{I}_i(n, a) = \mathsf{I}_i(n, a)$.

With this said, let us move to the proof theory, which is based on defeasible theories of the form

$$(F, R^C, R^I, \prec)$$

where R^C is as usual the set of constitutive rules, while R^I contains interpretation rules as illustrated above.

The peculiarity of proof theory—for the details see [Rotolo *et al.*, 2015]—is that we can distinguish the *derivation of interpretative acts* (A-interpretation) such as $\mathsf{I}_i(n, a)$—i.e., the fact that a certain I_i is obligatory or permitted for a certain legal provision n with outcome a—and the derivation of a certain interpretative outcome a as obligatory or permitted. ([Rotolo *et al.*, 2015] use the term *admissible* rather than *permitted*. For uniformity with other sections in this chapter we use the latter expression.)

As for *A-interpretation*, there are two ways to prove that an interpretation ϕ is permitted ($+\partial_{\mathsf{P}}\phi$, i.e., $\mathsf{P}\phi$): (1) $\mathsf{P}\phi$ or $\mathsf{O}\phi$ are a fact, or (2) $\mathsf{P}\phi$ must be derived by the rules of the theory. In the second case, three conditions must hold: (2.1) any complementary of $\mathsf{P}\phi$ does belong to the facts; (2.2) there must be a rule introducing the permissibility for ϕ

which can apply; (2.3) every rule s for $\sim\phi$ is either discarded or defeated by a stronger rule for ϕ.

The proof conditions for obligatory A-interpretation ($+\partial_{\mathsf{O}}\phi$, i.e., $\mathsf{O}\phi$) are much easier but we need to work on the fact that ϕ is an interpretation of any given provision n and we have to make explicit its structure. Indeed, that an interpretation I_i for the provision n is obligatory means that I_i is permitted and that no other (non-conflicting) interpretations for n is permitted.

As for *O-interpretation*, l is derivable as permitted if there is a permitted A-interpretation for any provision n resulting in l (i.e., $\mathsf{PI}_i(n, l)$); if any alternative A-interpretation for n is refuted, then l is derived as obligatory.

Let us illustrate this intuition with an example.

Example 12.2. *Consider art. 575 of the Italian penal code*

> *Art. 575. Homicide. Whoever causes the death of a* man *[*uomo*] is punishable by no less than 21 years in prison.*

art. 578 of the Italian penal code

> *Art. 578. Infanticide. The mother who causes the death of her newborn baby immediately after birth [. . .], when this act is committed in conditions of material or moral distress, is punishable with a sentence between 4 and 12 years of prison.*

and art. 3 of the Italian constitution (see above). Assume that

a = *Whoever causes the death of a adult male person is punishable by no less than 21 years in prison*
a' = *Whoever causes the death of a male person is punishable by no less than 21 years in prison*
b = *Whoever causes the death of a person is punishable by no less than 21 years in prison*
c = *All persons have equal social status and are equal before the law, without regard to their sex, race, language, religion, political opinions, and personal or social conditions.*

I_1 = *Literal interpretation or from ordinary meaning*
I_2 = *Interpretation from general principles*
I_3 = *Interpretation from substantive reasons*
I_4 = *Interpretation by coherence*

The following theory reconstructs an interpretive toy scenario in the Italian legal system.

$$
\begin{aligned}
F = \{&kill\_adult, kill\_female, \mathsf{O}\,\mathsf{I}_3(art.3, c), \mathsf{O}\,\mathsf{I}_1(art.578, d)\}\\
R = \{&r_1 : kill\_adult, kill\_female, \mathsf{O}\,\mathsf{I}_3(art.3, c) \Rightarrow^I\\
&\qquad \mathsf{I}_2(art.575, b) \otimes \mathsf{I}_3(art.575, b) \otimes \neg \mathsf{I}_1(art.575, a),\\
&r_2 : \mathsf{O}\,\mathsf{I}_1(art.578, d) \Rightarrow^I \mathsf{I}_1(art.575, a) \otimes \mathsf{I}_4(art.575, a),\\
&r_3 :\Rightarrow^I \mathsf{I}_1(art.575, a'),\\
&r_4 : b \Rightarrow_C a\}\\
\prec = \{&r_2 \prec r_1\}
\end{aligned}
$$

Rule r_1 says that, if art. 3 of the Italian constitution states formal equality before the law without regard also to gender identity, then b is the best interpretation outcome of art. 575 of the penal code, with I_2 (e.g., interpretation from general principles) as preferred over I_3 (say, interpretation from substantive reasons). In the light of art. 3, if these two interpretive options are refuted by any stronger interpretive conclusions, then the last sub-ideal option is to reject the interpretation I_1 from ordinary meaning (according to which only the homicide of adult male persons is punishable!). Rule r_2 states that, in case one kills an adult person and art. 578 ought to be interpreted literally, then we have a reason to interpret art. 575 by coherence (with respect to art. 578) as a. Rule r_3 establishes by default that art. 575 be literally interpreted as a'. Finally, rule r_4 is a constitutive rule saying that b entails a, i.e., that a provision punishing whoever causes the death of a person entails that a provision should be in the system that punishes whoever causes the death of a an adult male person.

What can we derive as A-interpretations*? Facts make rules r_1 and r_2 applicable. Rule r_3 has an empty antecedent, so it is applicable, too. Despite $r_4 \in R$, r_1 and r_2 are in conflict. The theory assumes that r_2 is stronger than r_1, thus we would obtain $+\partial^I_{\mathsf{P}}\mathsf{I}_1(art.575, a)$ (and so $-\partial^I_{\mathsf{P}}\mathsf{I}_2(art.575, b)$ and $-\partial^I_{\mathsf{P}}\mathsf{I}_3(art.575, b)$). However, these last conclusions are not obtained because of r_3: r_3 and r_2 attack each other, thus we in fact have $-\partial^I_{\mathsf{P}}\mathsf{I}_1(art.575, a)$, $-\partial^I_{\mathsf{P}}\mathsf{I}_1(art.575, a')$, and also $-\partial^I_{\mathsf{P}}\mathsf{I}_4(art.575, a)$ (i.e., they are all refuted). Hence, we reinstate $+\partial^I_{\mathsf{P}}\mathsf{I}_2(art.575, b)$ via r_1. What interpretations are obligatory? Trivially, we get $+\partial^I_{\mathsf{O}}\mathsf{I}_3(art.3, c)$. Also, since $\mathsf{I}_2(art.575, b)$ is the only admissible interpretation of art. 575, then $+\partial^I_{\mathsf{O}}\mathsf{I}_2(art.575, b)$. All other interpretations are refuted as obligatory.*

As for O-interpretation*, we can only show that $+\partial^{\Box} b$, where $\Box \in \{\mathsf{P}, \mathsf{O}\}$.*

12.3 Further readings

All technical details of the logic, including complexity results are offered in [Rotolo *et al.*, 2015].

In that work, a second variant of the logic is introduced that models interpretation when provisions are logically *structured*, i.e., when any n corresponds to a linguistic sentence having the structure of a rule like $a_1, \ldots, a_n \Rightarrow_{\mathsf{O}} b$: this means that n is semi-interpreted provision, since expressing the logical structure of n requires an interpretive effort on the original textual version of n. Interpreting n amounts to considering the components $a_1, \ldots, a_n, b$ of n and ascribing to them a meaning. In other words, given a deontic rule $r : a_1, \ldots, a_n \Rightarrow_{\mathsf{O}} b$, an interpretation function maps the sequence $x = \langle a_1, \ldots, a_n, b\rangle$ of literals in r onto another sequence y of literals that can be identical (literal interpretation), partially different or completely different from x. Hence, an interpretation I_i is meant to make the original version of rule r unusable and the new one—where the literals are changed according to y—usable to derive an obligation. For instance, if $r : a_1, a_2 \Rightarrow_{\mathsf{O}} b$ and the interpretation I_i returns $y = \langle a_1, a_2', b'\rangle$, the *interpreted version* of r according to I_i is $r : a_1, a_2' \Rightarrow_{\mathsf{O}} b'$.

A further extension for handling interpretation across different legal systems—something happening, e.g., in private international law—is presented in [Malerba *et al.*, 2016].

13 Defeasible deontic logic with time

13.1 Basics

The extension of Defeasible Logic with deontic operators makes the the logic more expressive and more capable of representing aspects of legal reasoning insofar as it allows us to consider the important distinction between constitutive rules and prescriptive rules, and to differentiate among normative effects. However, a key element is still missing: time. Very often norms have temporal parameters and Deontic Defeasible Logic is not able to reason about them. In this section we are going to extend the logic with temporal parameters. In particular we are going to *temporalise* the logic. This means that we attach a temporal parameter to the atomic elements of the logic, i.e., to the atomic propositions. For the logic we assume a discrete totally ordered set of instants of time $\mathcal{T} = \{t_0, t_1, t_2, \ldots\}$. Based on this we can introduce the notion of *temporalised literals*. Thus if l is a plain literal, i.e., $l \in$ PlainLit, and

$t \in \mathcal{T}$ then l^t is a temporalised literals. The intuitive interpretation of l^t is that l is true (or holds) at time t. We use TempLit to denote the set of temporalised literals. Deontic literals are now obtained from temporalised literals using the same conditions as in Section 9; thus a deontic literal is an expression like $\mathsf{O}l^t$, where its natural reading is that l is obligatory at time t, or that the obligation of l is in force at time t. Finally, given a time instant t and $y \in \{pers, tran\}$ we call the combination of (t, y) *duration specification*, and literals labelled with a duration specification duration literals. A duration literal has the form $l^{(t,y)}$. We denote the set of duration literals DurLit. The set of literals is now composed by the set of temporalised literals and the set of deontic literals, namely Lit = DeonLit ∪ TempLit. The signature of rules is now

$$\text{Rule}\colon 2^{\text{Lit}} \times \text{DurLit} \tag{27}$$

this means that a rule has the following form

$$r\colon a_1^{t_1}, \ldots, a_n^{t_n} \hookrightarrow_X c^{(t,y)} \tag{28}$$

where $X \in \{C, \mathsf{O}\}$, specifying whether the rules is a constitutive or a prescriptive one, and $y \in \{tran, pers\}$ indicating whether the conclusion of the rule is either *transient* or *persistent*.

The idea behind the distinction between a transient and persistent conclusion is whether the conclusion is guaranteed to hold for a single instant or it continues to hold until it is terminated. This is particular relevant for prescriptive rules, since their conclusions are obligations (or, in general deontic effects), and obligations, once triggered, remain in force until they are complied with, violated, or explicitly terminated. Accordingly we can use the duration specification $(t, tran)$ to indicate that on obligation is in force at a specific time t, and must be fulfilled at that time, while the duration specification $(t, pers)$ establishes that an obligation enters in force at time t.

The inference mechanism extends that of Defeasible Deontic Logic taking into account the temporal and durations specification. To assert that p holds at time t we have two ways:

1. Give an argument for p at time t';[48]
2. Evaluate all counterarguments against it. Here, we have a few cases:

[48] We equate arguments with rules, thus this is the same as saying that there is a (defeasible) rule such that all the elements in its antecedent are provable and the conclusion is $p^{(t',y)}$.

(a) If the duration specification of p is $(t, tran)$ $(t' = t)$, then, the counterargument must be for the same time t given that p is ensured to hold only for t.

(b) If the duration specification of p is $(t', pers)$, then t' can precede t and we can 'carry' over the conclusion from previous times. In this case, the counterarguments we have to consider are all rules whose conclusion has a duration specification (t'', z) such that $t' \leq t'' \leq t$.

3. Rebut the counterarguments. This is the same as the corresponding step of basic defeasible logic, the only thing to pay attention to is that when we rebut with a stronger argument, the stronger argument should have t'' in the duration specification of the conclusion.

The general idea of the conditions outline above is that, as we have already alluded to, it is possible to assert that something holds at time t, because it did hold at time t', $t' < t$, by persistence, but there must be no reasons to terminate it. Thus new information defeats previous one.

To illustrate the intuition we just described consider Section 8.2.1.a of the Australian Telecommunications Consumers Protection Code 2012 (TCPC 2012).

> A Supplier must take the following actions to enable this outcome:
>
> (a) **Demonstrate fairness, courtesy, objectivity and efficiency:** Suppliers must demonstrate, fairness and courtesy, objectivity, and efficiency by:
>
> (i) Acknowledging a Complaint:
>
> A. immediately where the Complaint is made in person or by telephone;
>
> B. within 2 Working Days of receipt where the Complaint is made by email; ....

The normative fragment above can be represented by the following set of rules:

$$tcpc_1 : Complaint^t, inPerson^t \Rightarrow_O Acknowledge^{(t,tran)}$$
$$tcpc_2 : Complaint^t \Rightarrow_O Acknowledge^{(t,pers)}$$
$$tcpc_3 : Complaint^t \rightsquigarrow_O \neg Acknowledge^{(t+2d,tran)}$$

Rule $tcpc_1$ covers the case of a complaint made in person of by phone. Given that the complaint must be acknowledged immediately, we can

use the duration specification $(t, tran)$, where t is the time when the complaint is received. The *tran* specification implies that the obligation to acknowledge the complain is in force only at t and not acknowledging at t results in a violation. For the case regulated by paragraph B, we use two rules. The first $tcpc_2$ is to initiate the obligation (at the same time t when the complaint is received, while $tcpc_3$ gives the deadline by when the content of the obligation must be fulfilled. Notice that we use a defeater to terminate the obligation.

Suppose we have a complaint by email on day 10. From this we can derive $+\partial_{\mathsf{O}} Acknowledge^{10}$ from rule $tcpc_2$. By persistence we have that $+\partial_{\mathsf{O}} Acknowledge^{11}$. On day 12 the effect of rule $tcpc_3$ kicks in, and we have $-\partial_{\mathsf{O}} Acknowledge^{12}$.

Different versions of Temporal Defeasible Logic have been developed: we refer the reader to [Governatori *et al.*, 2005b; Governatori *et al.*, 2007a; Governatori and Rotolo, 2013].

13.2 From rules to meta-rules

The temporal Defeasible Logic just presented allows us to reason about the times specified inside norms, but it is not able to capture the lifecycle of norms. To obviate this problem [Governatori and Rotolo, 2010] propose to consider a legal system as a time-series of its versions, where each version is obtained from previous versions by some norm changes, e.g., norms entering in the legal system, modification of existing norms, repeals of existing norms, ... This means that we can represent a legal system LS as a sequence

$$LS(t_1), LS(t_2), \ldots, LS(t_j) \tag{29}$$

where each $LS(t_i)$ is the snapshot of the rules (norms) in the legal system at time t_i. Graphically it can be represented by the picture in Figure 1

A *rule* is a relation between a set of premises (conditions of applicability of the rule) and a conclusion. In this paper the admissible conclusions are either literals or rules themselves; in addition the conclusions and the premises will be qualified with the time when they hold. We consider two classes of rules: *meta-rules* and *proper rules*. Meta-rules describe the inference mechanism of the institution on which norms are formalised and can be used to establish conditions for the creation and modification of other rules or norms, while proper rules correspond to norms in a normative system. In what follows we will use Rule to denote the set of rules, and MetaRules for the set of meta-rules, i.e., rules whose consequent is a rule.

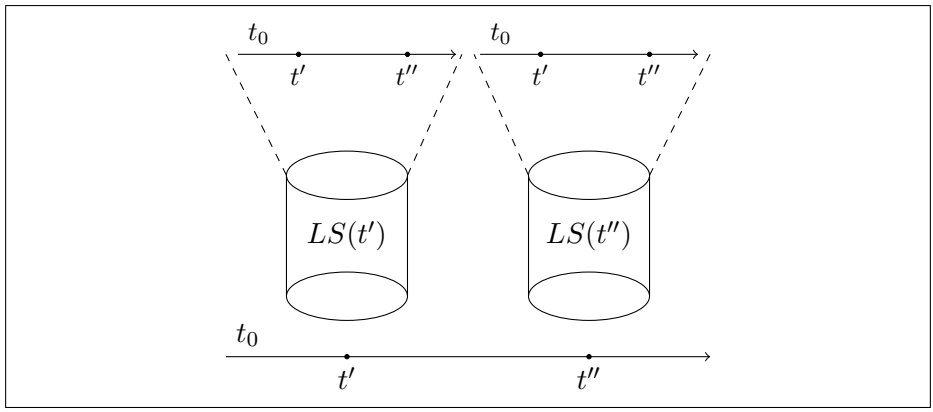


Figure 1: Legal System at t' and t''

A *temporalised rule* is either an expression $(r\colon \bot)^{(t,x)}$ (the void rule) or $(r\colon \emptyset)^{(t,x)}$ (the empty rule) or $(r\colon A \hookrightarrow_X B)^{(t,x)}$, where r is a rule label, A is a (possibly empty) set of temporalised literals, $X \in \{C, \mathsf{O}\}$, B is a duration literal, $t \in \mathcal{T}$ and $x \in \{tran, pers\}$.

We have to consider two temporal dimensions for norms in a normative system. The first dimension is when the norm is in force in a normative system, and the second is when the norm exists in the normative system from a certain viewpoint. So far temporalised rules capture only one dimension, the time of force. To cover the other dimension we introduce the notion of temporalised rule with viewpoint. A *temporalised rule with viewpoint* is an expression

$$(r\colon A \hookrightarrow_X B)^{(t,x)}@(t', y), \tag{30}$$

where $(r\colon A \hookrightarrow_X B)^{(t,x)}$ is a temporalised rule, $t' \in \mathcal{T}$ and $y \in \{tran, pers\}$.

Finally, we introduce meta-rules, that is, rules where the conclusion is not a simple duration literal but a temporalised rule. Thus a *meta-rule* is an expression

$$(s\colon A \hookrightarrow (r\colon B \hookrightarrow_X C)^{(t',x)})@(t, y), \tag{31}$$

where $(r\colon B \hookrightarrow_X C)^{(t',x)}$ is a temporalised rule, $r \neq s$, $t \in \mathcal{T}$ and $y \in \{tran, pers\}$. Notice that meta-rules carry only the viewpoint time (the validity time) but not the "in force" time. The intuition behind this is that meta-rules yield the conditions to modify a legal system. Thus they specify what rules (norms) are in a normative system, at what time the rules are valid, and the content of the rules. Accordingly, these rules

must have an indication when they have been inserted in a normative system, but then they are universal (i.e., apply to all instants) within a particular instance of a normative system.

Every temporalised rule is identified by its rule label and its time. Formally we can express this relationship by establishing that every rule label r is a function

$$r\colon \mathcal{T} \mapsto \text{Rule}. \tag{32}$$

Thus a temporalised rule r^t returns the value/content of the rule 'r' at time t. This construction allows us to uniquely identify rules by their labels[49], and to replace rules by their labels when rules occur inside other rules. In addition there is no risk that a rule includes its label in itself. In the same way a temporalised rule is a function from $\mathcal{T}$ to Rule, we will understand a temporalised rule with viewpoint as a function with the following signature:

$$\mathcal{T} \mapsto (\mathcal{T} \mapsto \text{Rule}). \tag{33}$$

As we have seen above a legal system LS is a sequence of versions $LS(t_0), LS(t_1), \ldots$. The temporal dimension of viewpoint corresponds to a version while the temporal dimension temporalising a rule corresponds to the time-line inside a version. Thus the meaning of an expression $r^{t_v}@t_r$ is that we take the value of the temporalised rule r^{t_v} in $LS(t_r)$. Accordingly, a version of LS is just a repository (set) of norms (implemented as temporal functions).

Accordingly, given a rule r, the expression $r^t@t'$ gives the value of the rule (set of premises and conclusion of the rule) at time t in the repository t'. The content of a void rule, e.g., $(r\colon \bot)^t@t'$ is $\bot$, while for the empty rule the value is the empty set. This means that the void rule has a value for the combination of the temporal parameters, while for the empty rule, the content of the rule does not exist for the given temporal parameters. Another way to look at the difference between the empty rule and the void rule is to consider that a rule is a relationship between a set of premises and a conclusion. For the void rule this relationship is between the empty set of premises and the empty conclusion; thus the rule exists but it does not produce any conclusion. For the empty rule, the relationship is empty, thus there is no rule. Alternatively, we can think of the function corresponding to temporalised rules as a partial

[49] We do not need to impose that the function is injective: while each label should have only one content at any given time, we may have that different labels (rules) have the same content.

function, and the empty rule identifies instants when the rule is not defined.

For a transient fully temporalised literal $l^{(t,x)}@(t', tran)$ the reading is that the validity of l at t is specific to the legal system corresponding to repository associated to t', while $l^{(t,x)}@(t', pers)$ indicates that the validity of l at t is preserved when we move to legal systems after the legal system identified by t'. An expression $r^{(t,tran)}$ sets the value of r at time t and just at that time, while $r^{(t,pers)}$ sets the values of r to a particular instance for all times after t (t included).

We will often identify rules with their labels, and, when unnecessary, we will drop the labels of rules inside meta-rules. Similarly, to simplify the presentation and when possible, we will only include the specification whether an element is persistent or transient only for the elements for which it is relevant for the discussion at hand.

Meta-rules describe the inference mechanism of the institution on which norms are formalised and can be used to establish conditions for the creation and modification of other rules or norms, while proper rules correspond to norms in a normative system. Thus a temporalised rule r^t gives the 'content' of the rule 'r' at time t; in legal terms it tells us that norm r is in force at time t. The expression

$$(p^{t_p}, q^{t_q} \Rightarrow (p^{t_p} \Rightarrow_{\mathsf{O}} s^{(t_s,pers)})^{(t_r,pers)})@(t, tran) \tag{34}$$

means that, for the repository at t, if p is true at time t_p and q at time t_q, then $p^{t_p} \Rightarrow_{\mathsf{O}} s^{(t_s,pers)}$ is in force from time t_r onwards.

A legal system is represented by a temporalised defeasible theory, called *normative theory*, i.e., a structure

$$(F, R, R^{\text{meta}}, \prec) \tag{35}$$

where F is a finite set of facts (i.e., fully temporalised literals), R is a finite set of prescriptive and constitutive rules, R^{meta} is a finite set of meta rules, and $\prec$, the superiority relation over rules is formally defined as $\mathcal{T} \mapsto (\mathcal{T} \mapsto \text{Rule} \times \text{Rule})$. accounting that we can have different instances of the superiority relation depending on the legal systems (external time) and the time when the rules involved in the superiority are evaluated[50].

In the current logic a conclusion has a form like: $+\partial t@t'\ p^{t_p}$, meaning that the conclusion that p holds at time t_p is derivable at time t using the information included in the version of the legal system at time t'.

[50]For instance, if we have $s \prec^{2007}_{Monday} r$ and $r \prec^{2007}_{Tuesday} s$, it means that, according to the regulation in force in 2007, on Monday rule s is stronger than rule r, but on Tuesday r is stronger than s.

The inference mechanism with meta-rules is essentially an extension of that of temporal defeasible logic, but it involves more steps. Rules are no longer just given, but they can be derived from meta-rules. Thus to prove $+\partial t@t' p^{t_p}$ the first thing to do is to see if it is possible to derive a rule r having p^{t_p} as its head. But we have to derive such rule at the appropriate time. Here, we want to remember that a rule is a function from time (validity time or version of a legal system) to time (when a rule is in force in a version of a legal system) to the content of the rule (relationship between a set of premises and a conclusion). The basic intuition is that a rule corresponds to a norm, and there could be several modifications of a norm, thus deriving a rule means to derive one of such modifications. As we shall see in the next section a meta-rule (or more generally a set of meta-rules) can be used to encode a modification of a norm. In general it is possible to have multiple (conflicting) modifications of a norm. Accordingly, to derive a rule, we have to check that there are no conflicting modifications[51] or the conflicting modifications are weaker than the current modification. The final consideration is that in this case we have two temporal dimensions, and the persistence applies to both. Thus we can have persistence inside a legal system, thus we can conclude $+\partial t''@t'\ p^{t_p}$ from $+\partial t@t'\ p^{t_p}$, where $t < t''$ as well as persistence over versions, thus $+\partial t@t''\ p^{t_p}$ from $+\partial t@t'\ p^{t_p}$, where $t' < t''$.

Sections 13.1 and 13.2 were meant to provide the full logical machinery for modelling legal dynamics. Section 14 will extensively illustrate with several realistic examples how to apply such a machinery in the legal domain.

13.3 Further readings

The most comprehensive version of Temporal Defeasible Logic has been presented in [Governatori and Rotolo, 2010] where rules and meta-rules are used. For thorough presentations of Temporal Defeasible Logic (without meta-rules), its properties and application to modelling obligation with time and deadlines we refer the reader also to [Governatori *et al.*, 2005b; Governatori *et al.*, 2007a; Governatori and Rotolo, 2013]. In particular, [Governatori *et al.*, 2005b] was the first paper proposing a temporal extension of the basic logic for modelling normative positions, [Governatori *et al.*, 2007a] developed a version with time intervals, while [Governatori and Rotolo, 2013] proved that the complexity of the logic is still linear.

[51] Two meta-rules are conflicting, when the two meta-rules have the same rule as their head, but with a different content.

14 Modelling legal changes

Many formalisms have been proposed in the literature with the purpose of modelling legal dynamics. An overview was offered in Section 5.4. This section shows how the Defeasible Deontic Logic with Time presented in the previous section can be used for this purpose.

14.1 Types of legal change

Norm changes in the law are performed by norms affecting the legal system, and can be explicit or implicit [Governatori *et al.*, 2005a; Governatori *et al.*, 2007b; Governatori and Rotolo, 2010][52]:

Explicit: The law introduces norms whose peculiar objective is to change the system by specifying what and how other existing norms should be modified;

Implicit: the legal system is revised by introducing new norms which are not specifically meant to modify previous norms, but which change in fact the system because they are incompatible with such existing norms and prevail over them. (The new norms prevail because, for example, have a higher ranking status in the hierarchy of the legal sources or because have been subsequently enacted.)

The most interesting case is when we deal with explicit modifications, which permit to classify a large number of modification types (and which include those that we can implicitly apply): indeed, when modifications are implicit, we can handle conflicts using standard criteria in defeasible reasoning for conflict-detection and -resolution. In general, we have different types of modifying norms, as their effects (the resulting modifications) may concern, for example, the text of legal provisions, their scope, or their time of force, efficacy, or applicability, or their own existence or validity [Guastini, 1998; Governatori *et al.*, 2005a; Governatori *et al.*, 2007b].

Derogation is an example of scope change: a norm n supporting a conclusion p and holding at the national level may be derogated by a norm n' supporting a different conclusion p' within a regional context. Hence, derogation corresponds to introducing one or more exceptions to n. Temporal changes impact on the target norm in regard to its date of force (the time when the norm is "usable"), date of effectiveness (when the norm

[52] Accordingly, a temporary norm n is not an example of norm change, since the termination of its validity, due to its temporariness, does not depend on any other norm affecting n.

in fact produces its legal effects) or date of application (when conditions of norm applicability hold). An example of change impacting on time of force is when a norm n is originally in force in 2007 but a modification postpones n to 2008. Substitution is an example of textual modification, as it generically replaces some textual components of a provision with other components. For instance, some of its applicability conditions are replaced by other conditions. Finally, we have modifications on norm validity and existence, such as abrogation and annulment. For instance, an annulment is usually seen as a kind of repeal, as it makes a norm invalid and removes it from the legal system. As we will see, its peculiar effect applies *ex tunc*: annulled norms are prevented to produce all their legal effects, independently of when they are obtained.

The following subsection illustrates the above types of modifications using real-life examples. However, in order to keep the presentation simple, we model simple scenarios using a few meta-rules.[53]

14.2 Modifications of scope: Derogation

Derogations are modifications of norm scope. Consider the following example based on art. 3 of the Italian Constitution (enacted in 1948):

Example 14.1 (Derogation)**.**

> *[***Target of the modification***] Article 3 (1) All citizens have equal social status and are equal before the law, without regard to their sex, race, language, religion, political opinions, and personal or social conditions.*

A example of (fictional) derogation is the following:

> *[***Modification enacted in 2014 and effective in 2015***] In derogation to the provisions set out in Article 3, paragraph 1, of the Constitution, the citizens who are resident in Bologna may have different social status, but this modification will be effective only in 2015, when Italy will be no longer in EU.*

From the logical point of view, derogation can be simply modeled by adding exceptions, in particular defeaters. Using meta-rules, Example

[53] Clearly, the utility of the machinery is evident when we have complex cases where more modifications and meta-rules interplay together and defeasible conditions for the modifications are considered in the reasoning process. However, even with simple cases we have discussed in Section 5.4 general and conceptual reasons why formalisms like this one is needed.

14.1 can be captured as follows[54].

Example 14.2 (Derogation (cont'd))**.** *Let* $D = (F, R, R^{meta}, \prec)$ *be a normative theory such that*

$$\mathit{Art.\ 3} : (\mathit{Citizen}^x \Rightarrow_{\mathsf{O}} \mathit{EqualStatus}^x)^{(1948,pers)}@(1948, tran) \in R$$

Example 14.1 is modeled by stating that R^{meta} *includes the following meta-rule*

$$\begin{array}{l} \mathit{derog}_{\mathit{Art.\ 3}} : (\sim EU^x \Rightarrow \\ \quad (r' : \mathit{Citizen}^x, \mathit{ResidentBologna}^x \rightsquigarrow_{\mathsf{O}} \sim \mathit{EqualStatus}^x)^{(2015,pers)})@(2014, pers) \end{array}$$

and that $\prec$ *is as follows (where* $t \geq 2015$*)*[55]*:*

$$\begin{array}{l} \{s \prec^{2014}_{2015} r' : s \in R[\mathit{EqualStatus}^x] \text{ and } A(s) \cap \partial^-(D) \neq \emptyset\} \in \prec \\ \{mr \prec^{2014}_{2015} \mathit{derog}_{\mathit{Art.\ 3}} : mr \in R^{meta}[\sim r^t] \text{ and } A(mr) \cap \partial^-(D) \neq \emptyset\} \in \prec \end{array}$$

Notice that the above conditions on $\prec$ ensures that this operation minimises the impact of the added meta-rule and the related defeater. In fact, the operation works on art. 3 (and any other similar provision) only when any conflicting meta-rule and art. 3 are applicable.

14.3 Textual modifications: Substitution

Consider a textual modification such as substitution, which typically replaces some textual components of a provision with other textual components. Another fictional (but this time reasonable!) example from the Italian constitution is the following:

Example 14.3 (Substitution)**.**

> *[***Target of the modification***] Article 3 (1) All citizens have equal social status and are equal before the law, without regard to their sex, race, language, religion, political opinions, and personal or social conditions.*
>
> *[***Modification enacted and effective in 2014***] In the Article 3, paragraph 1 of the Italian constitution the expression "citizens" is replaced with "human beings".*

[54]In the remainder of the paper, when temporal parameters are not essential we will not specify them and will just add a superscript x.

[55]Recall that, for any rule s, $A(s)$ denotes the set of antecedents of s, while $\partial^-(D)$ stands for the set of negative conclusions of the theory D. i.e., the literals occurring in conclusions of the form $-\partial$.

This example can be represented as follows.

Example 14.4 (Substitution (cont'd))**.** *Let* $D = (F, R, R^{meta}, \prec)$ *be a normative theory such that*

$$\mathit{Art.\ 3} : (\mathit{Citizen}^x \Rightarrow_{\mathsf{O}} \mathit{EqualStatus}^x)^{(1948,pers)}@(1948, tran) \in R,$$

where the substitution is modelled by the following meta-rule in R^{meta}

$$sub_{\mathit{Art.\ 3}} : (\Rightarrow (\mathit{Art.\ 3} : \mathit{HumanBeing}^x \Rightarrow_{\mathsf{O}} \mathit{EqualStatus}^x)^{(2014,pers)})@(2014, pers)$$

and $\prec$ *is as follows (where* $t \geq 2014$*):*

$\{s \prec^{2014}_{2014} \mathit{Art.\ 3}^{2014} : s \in R[\mathit{EqualStatus}^x]$ *and* $A(s) \cap \partial^-(D) \neq \emptyset\} \in \prec$
$\{sub_{\mathit{Art.\ 3}} \prec^{2014}_{2014} \mathit{Art.\ 3}^{2013}\} \in \prec$
$\{mr \prec^{2014}_{t} sub^{2014}_{\mathit{Art.\ 3}}, mr \in R^{meta}[\sim \mathit{Art.\ 3}^{2014}]$ *and*
$A(mr) \cap \partial^-(D) \neq \emptyset\} \in \prec$.

14.4 Temporal modifications

Temporal modifications are performed by meta-rules that change norms in regard to their time of force, efficacy, or applicability. Consider this example:

Example 14.5 (Temporal modification)**.**

> *[***Target of the modification***] Legislative Act n. 124, 23 July 2008.*
> *[. . .]*
> *Art. 8. This legislative act is in force since the date of publication of the* Gazzetta Ufficiale *[23 August 2008]*
>
> *[***Modification enacted and effective at 1 August 2008***] Legislative Act n. 124, 23 July 2008 is in force since 1 January 2009.*

Example 14.5 is reconstructed as follows.

Example 14.6 (Temporal modification (cont'd))**.** *For the sake of simplicity, assume that the content of Legislative Act n. 124 is* $a^x_1, \ldots a^x_n \Rightarrow_{\mathsf{O}} b^x$*. Hence, we have that* R^{meta} *contains the following meta-rule modeling the enactment of Legislative Act n. 124;*

$$mr\colon (\Rightarrow (L.\ 124 : a^x_1, \ldots a^x_n \Rightarrow_{\mathsf{O}} b^x)^{(23\ August\ 2008,pers)})@(23\ July\ 2008, pers).$$

The modification at hand is expressed by having in R^{meta} *other two meta-rules* mr' *and* mr'' *such that*

$$temp'_{L.\ 124}\colon (\rightsquigarrow \sim(L.\ 124\colon a^x_1, \ldots a^x_n \Rightarrow_{\mathsf{O}} b^x)^{(23\ August\ 2008,pers)})@(1\ August\ 2008, pers)$$

$$temp''_{L.\ 124}\colon (\Rightarrow (L.\ 124\colon a^x_1, \ldots a^x_n \Rightarrow_{\mathsf{O}} b^x)^{(1\ January\ 2009,pers)})@(1\ August\ 2008, pers)$$

such that $(temp'_{L.\ 124} \prec^{23\ August\ 2008}_{1\ August\ 2008} mr) \in \prec$.

14.5 Modifications on norm validity and existence: Annulment vs. abrogation

The expression *repeal* is sometimes used to generically denote the operation of norm withdrawal. However, at least two forms of withdrawal are possible: annulment and abrogation.

An *annulment* makes the target norm invalid and removes it from the legal system. Its peculiar effect applies *ex tunc*: annulled norms are prevented to produce all their legal effects, independently of when they are obtained. Annulments typically operate when the grounds (another norm) for annulling are hierarchically higher in the legal system than the target norm which is annulled: consider when a legislative provision is annulled (typically by the Constitutional Court) because it violates the constitution.

An *abrogation* works differently; the main point is usually that abrogations operate *ex nunc* and so do not not cancel the effects that were obtained from the target norm before the modification. If so, it seems that abrogations cannot operate retroactively. In fact, if a norm n_1 is abrogated in 2012, its effects are no longer obtained after then. But, if a case should be decided at time 2013 but the facts of the case are dated 2011, n_1, if applicable, will anyway produce its effects because the facts held in 2011, when n_1 was still in force (and abrogations are not retroactive). Accordingly, n_1 is still in the legal system, even though is no longer in force after 2012. Abrogations typically operate when the grounds (another norm) for abrogating is placed at the same level in the hierarchy of legal sources of the target norm which is abrogated: consider when a legislative provision is abrogated by a subsequent legislative act.

Consider this case:

Example 14.7 (Abrogation vs. Annulment)**.**

> *[***Target of the modification***] Legislative Act n. 124, 23 July 2008*
> *Art. 1. With the exception of the cases mentioned under the Articles 90 and 96 of the Constitution, criminal proceedings against the President of the Republic, the President of the Senate, the President of the House of Representatives, and the Prime Minister, are suspended for the entire duration of tenure. [. . .]*

In case of abrogation, we could have that the legislator enacts the following provision:

> *[**Abrogation enacted and effective at 1 January 2011**]*
> *Legislative Act n. 124, 23 July 2008 is abrogated.*

In case of (judicial) annulment, we would rather have

> *[**Annulment enacted and effective at 1 January 2011**]*
> *On account of Art. 3 of the Constitution [...] the Constitutional Court hereby declares the constitutional illegitimacy of Art. 1 of the Act n. 124, 23 July 2008.*

As we have recalled, the difference between the two cases is that the annulment has retroactive effects. In particular, let us focus on the following provisions from the Italian penal code:

> Art. 157 Italian of Penal Code – Terms of statute-barred penal provisions.
> When the the terms for statute-barred penal effects expire, the corresponding crime is canceled [...]
> Art. 158 Italian Penal Code – Effectiveness of the terms of statute-barred penal provisions The effectiveness of terms of statute-barred penal provisions begins starting from the time when the crime was committed.
> Art. 159 Italian of Penal Code – Suspension of time limits for statute-barred penal effects. The terms for statute-barred penal effects [...] are suspended whenever the criminal proceedings are suspended under any legislative provisions [...]

Consider a hypothetical case where the Italian Prime Minister is accused in 2007 of accepting bribes at the beginning of 2006. Clearly, if Legislative Act n. 124 is abrogated in 2011, since abrogation has no retroactive effects, art. 159 of Italian Penal Code applied from 2008 to 2011, and so the counting of terms has been suspended between these two years. Hence, from the perspective of 2011 (immediately after the abrogation) the relevant time passed is two years and six months (2006, 2007, and until July 2008). Instead, if the act is annulled in 2011, more time has passed from the perspective of 2011, because it is as if the Legislative Act n. 124 were never enacted: from 2006 until 2011.

As we can see, modeling retroactive legal modifications is far from obvious. The logical model proposed in [Governatori and Rotolo, 2010] and recalled in Section 13 offers a solution. In the next section we will illustrate the intuition and apply to the above example of annulment and abrogation.

14.6 Intermezzo – temporal dynamics and retroactivity

As we have previously argued, if $t_0, t_1, \ldots, t_j$ are points in time, the dynamics of a legal system LS can be captured by a time-series $LS(t_0), LS(t_1), \ldots, LS(t_j)$ of its versions. Each version of LS is like a norm repository: the passage from one repository to another is effected by legal modifications or simply by temporal persistence. This model is suitable for modeling complex modifications such as retroactive changes, i.e., changes that affect the legal system with respect to legal effects which were also obtained before the legal change was done.

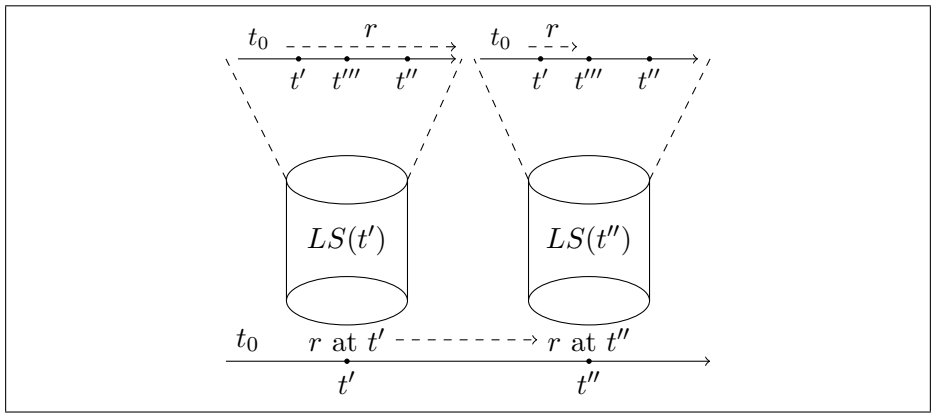
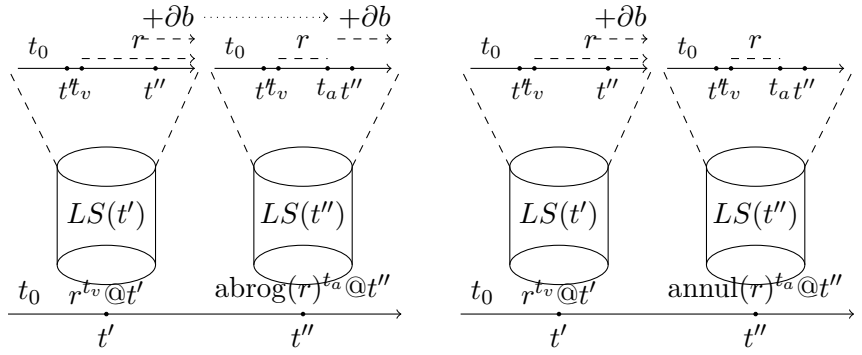


Figure 2: Legal System at t' and t''

The dynamics of norm change and retroactivity need to fully make use of the time-line within each version of LS (the time-line placed on top of each repository in Figure 2). Clearly, retroactivity does not imply that we can really change the past: this is "physically" impossible. Rather, we need to set a mechanism through which we are able to reason on the legal system from the viewpoint of its current version but *as if* it were revised in the past: when we change some $LS(i)$ retroactively, this does not mean that we modify some $LS(k)$, $k < i$, but that we move back from the perspective of $LS(i)$. Hence, we can "travel" to the past along this inner time-line, i.e., from the viewpoint of the current version of LS where we modify norms.

Figure 2 shows a case where the legal system LS and its norm r persist from time t' to time t'' and can have effects immediately from t'. Now, the figure represents the situation where r is retroactively repealed at t'' by stating that the modification applies from t''' (which is between t' and t'') onwards. The difference between abrogation and annulment is illustrated in Figures 3(a) and 3(b).



(a) Abrogation. In $LS(t')$ rule r produces a persistent effect b. Literal b carries over by persistence to $LS(t'')$ even if r is no longer in force.

(b) Annulment. In $LS(t')$ rule r is applied and produces a persistent effect b. Since r is annulled in $LS(t'')$, b must be undone as well.

Figure 3: Abrogation and Annulment

14.7 Modifications on norm validity and existence: Annulment vs. abrogation (cont'd)

On account of our previous considerations, the cases in Example 14.7 can be reconstructed as follows.

Example 14.8 (Abrogation vs Annulment (cont'd))**.** *First of all, for the sake of simplicity let us*

- *only consider the case of Prime Minister (Legislative Act n. 124 mentions other institutional roles),*
- *assume that the dates of enactment and effectiveness coincide and are generically 2008,*
- *the duration of tenure covers a time span from 2008 to 2012,*

and formalize the corresponding fragment of art. 1 of Legislative Act n. 124 (23 July 2008) as follows:

$$L.\ 124 : (Crime^{x}, Tenure^{x+y} \Rightarrow_{\mathsf{O}} Suspended^{(x+y,tran)})^{(2008,pers)})@(2008, pers)$$

The duration of tenure spanning from 2008 to 2012 is represented as follows:

$$r1 : (Elected^{2008} \Rightarrow_{\mathsf{O}} Tenure^{(2008,pers)})^{(2008,pers)})@(2008, pers)$$
$$r2 : (Elected^{2008} \rightsquigarrow_{\mathsf{O}} \neg Tenure^{2012})^{(2008,pers)})@(2008, pers)$$

Arts. 157-159 of the Italian Penal Code state the following:

$$\textit{Art. 157}\colon (\textit{Crime}^{x}, \textit{Terms}^{x+y} \Rightarrow_{\mathsf{O}} \textit{CrimeCancelled}^{(x+y,pers)})^{(z,pers)})@(z,pers)$$

$$\textit{Art. 158}\colon (\textit{Crime}^{x} \Rightarrow_{\mathsf{O}} \textit{Terms}^{(x,pers)})^{(z,pers)})@(z,pers)$$

$$\textit{Art. 159}\colon (\textit{Crime}^{x}, \textit{Suspended}^{x+y} \Rightarrow_{\mathsf{O}} \neg \textit{Terms}^{(x+y,tran)})^{(z,pers)})@(z,pers)$$

As proposed by [Governatori and Rotolo, 2010], the distinction between abrogation and annulment requires the distinguish between void *rules and* empty *rules. The content of a void rule, e.g.,* $(r\colon \bot)^{t}@t'$ *is* $\bot$*, while for the empty rule the value is the empty set. This means that the void rule has value for the combination of the temporal parameters, while for the empty rule, the content of the rule does not exist for the given temporal parameters.*

Given a rule $(r\colon A \Rightarrow b^{t_b})^{t_r}@t$*, the abrogation of* r *at* t_a *in repository* t' *is basically obtained by having in the theory the following meta-rule*

$$abr_r\colon \Rightarrow (r\colon \bot)^{(t_a,pers)})@(t', pers) \tag{36}$$

where $t' > t$*. The abrogation simply terminates the applicability of the rule. More precisely this operation sets the rule to the void rule. The rule is not removed from the system, but it has now a form where no longer can produce effects. In the case of the Legislative Act n. 124 (23 July 2008) we would have*

$$abr_{L.\ 124}\colon \Rightarrow (L.\ 124\colon \bot)^{(2011,pers)})@(2011, pers)$$

Hence, we can derive, for example

- $+\partial_{\mathsf{O}} x@x\ \textit{Suspended}^{x}$, $2008 \leq x \leq 2010$*;*
- $-\partial_{\mathsf{O}} x@x\ \textit{Terms}^{x}$, $2008 \leq x \leq 2010$*;*
- $-\partial_{\mathsf{O}} 2011@2011\ \textit{Suspended}^{2011}$*;*
- $+\partial_{\mathsf{O}} 2011@2011\ \textit{Terms}^{2011}$.

This is in contrast to what we do for annulment where the rule to be annulled is set to the empty rule. This essentially amounts to removing the rule from the repository. From the time of the annulment the rule has no longer any value. All past effects are thus blocked as well.

The definition of a modification function for annulment depends on the underlying variants of the logic, in particular whether conclusions persist across repositories. Minimally, the operation requires the introduction of a meta-rule setting the rule r *to be annulled to* $\emptyset$*, with the time when the rule is annulled and the time when the meta-rule is inserted in the legal system:*

$$(annul_r\colon \Rightarrow (r\colon \emptyset)^{(t_a,pers)})@(t', pers) \tag{37}$$

Hence,

$$(annul_{L.\,124}\colon \Rightarrow (L.\,124\colon \emptyset)^{(2008,pers)})@(2011, pers)$$

If we assume that conclusions persist over repositories we need some additional technical machinery to block pasts effects from previous repositories. In this case, since L. 124 *is modeled as a transient rule, we have basically to add a defeater like the following*[56]*:*

$$((annul_{ef}\colon \rightsquigarrow_{\mathsf{O}} \neg Suspended^{2008})^{(2008,pers)})@(2011, pers)$$

Hence, we now have, for example

- $-\partial_{\mathsf{O}} x@2011\ Suspended^{x},\ 2008 \leq x;$
- $+\partial_{\mathsf{O}} x@2011\ Terms^{x},\ 2008 \leq x.$

14.8 Further readings

The complete logic for legal modifications is offered in [Governatori and Rotolo, 2010] where two cases are extensively studied: abrogation and annulment. Preliminary versions covering more modifications are given by [Governatori *et al.*, 2005a; Governatori *et al.*, 2007b]. A recent work on temporary changes is [Cristani *et al.*, 2017].

15 Conclusion

The law is a complex phenomenon, which can be analysed into different branches according to the authority who produces legal norms and according to the circumstances and procedures under which norms are created. But, independently of these aspects, research in deontic logic has shown that it is possible to identify some formal features that legal norms and legal systems should enjoy.

We saw that, despite the variety of nor types, legal norms have often a conditional structure like [Kelsen, 1991; Sartor, 2005]

$$\text{if } A_1, \ldots, A_n \text{ then } B \tag{38}$$

where $A_1, \ldots, A_n$ are the applicability conditions of the norm and B denotes the legal effect which ought to follow when those applicability

[56] The general procedure to block conclusions when conclusions persist over repositories can be very complex: for all details, see [Governatori and Rotolo, 2010].

conditions hold[57].

Many aspects of norms and normative systems have been acknowledged in the field of legal theory and artificial intelligence and law, where there is now much agreement about the structure and properties of norms.

Definition 15.1. *Requirements for representing legal norms include the following:*

Norms vs Legal Provisions [Ross, 1958; Tarello, 1980]. *It is standard in legal theory to distinguish between legal provisions (authoritative legal texts) and legal norms, this last being the meaning of provisions resulting from the interpretive process.*

Norm properties [Gordon, 1995]. *Norms are objects with properties, such as*

Jurisdiction. *The limits within which the norm is authoritative and its effects are binding (of particular importance are spatial and geographical references to model jurisdiction).*

Authority [Prakken and Sergot, 1996]. *Who produced the norm, a feature which indicates the ranking status of the norm within the sources of law (whether the rule is a norm constitutional provision, a statute, is part of a contract clause or is the ruling of a precedent, and so on).*

Temporal properties [Governatori and Rotolo, 2010]. *Norms usually are qualified by temporal properties, such as:*

1. *the time when the legal provision and the corresponding norm is in force and/or has been enacted;*
2. *the time when the norm can produce legal effects (when the norm is applicable and supports the derivation of legal effects);*
3. *the time when the normative effects hold.*

[57]Indeed, norms can be also unconditioned, that is their effects may not depend upon any antecedent condition. Consider, for example, the norm "everyone has the right to express his or her opinion". Usually, however, norms are conditioned. In addition, unconditioned norms can formally be reconstructed in terms of (38) with no antecedent conditions.

Defeasibility [Gordon, 1995; Prakken and Sergot, 1996; Sartor, 2005]. *When the antecedent of a norm is satisfied by the facts of a case, the conclusion of the norm presumably holds, but is not necessarily true. Three types of legal defeasibility can be identified:*

Inference-based defeasibility: *It covers the fact that legal conclusions, though correctly supported by certain pieces of information, cannot be derived when the legal knowledge base including those information is expanded with further pieces of information.*

Process-based defeasibility: *It addresses the dynamic aspects of defeasible reasoning. As for legal reasoning, a crucial observation here is that it often proceeds according to the norms of legal procedures, such as those regulating the allocation of the burden of proof.*

Theory-based defeasibility: *It regards the evaluation and the choice of theories which explain and systematise the available legal input information (such as a set of precedents): when a better theory becomes available, inferior theories are to be abandoned.*

The defeasibility of legal norms gives rise to the above types of defeasibility and breaks down into the following issues:

Conflicts [Prakken and Sergot, 1996]. *Norms can conflict, namely, they may lead to incompatible legal effects. Conceptually, conflicts can be of different types, according to whether two conflicting norms*

- *are such that one is an exception of the other (i.e., one is more specific than the other);*
- *have a different ranking status;*
- *have been enacted at different times;*

Accordingly, norm conflicts can be resolved using principles about norm priorities, such as:

- lex specialis*, which gives priority to the more specific norms (the exceptions);*
- lex superior*, which gives priority to the norm from the higher authority (see 'Authority' above);*
- lex posterior*, which gives priority to the norm enacted later (see 'Temporal properties' above).*

Norm validity [Governatori and Rotolo, 2010]. *Norms can be invalid or become invalid. Deleting invalid norms is not an option when it is necessary to reason retroactively with norms which were valid at various times over a course of events. For instance:*

1. *The* annulment *of a norm is usually seen as a kind of repeal which invalidates the norm and removes it from the legal system as if it had never been enacted. The effect of an annulment applies* ex tunc*: annulled norms are prevented from producing any legal effects, also for past events.*
2. *An* abrogation *on the other hand operates* ex nunc*: The norm continues to apply for events which occurred before the norm was abrogated.*

Legal procedures. *Norms not only regulate the procedures for resolving legal conflicts (see above), but also for arguing or reasoning about whether or not some action or state complies with other, substantive norms [Governatori, 2005]. In particular, norms are required for procedures which*

1. *regulate methods for detecting violations of the law and checking compliance;*
2. *determine the normative effects triggered by norm violations, such as reparative obligations, namely, which are meant to repair or compensate violations*[58]*.*

Normative effects. *There are many normative effects that follow from applying norms, such as obligations, permissions, prohibitions and also more articulated effects such as those introduced, e.g., by Hohfeld [Sartor, 2005].*

Persistence of normative effects [Governatori *et al.*, 2005b]. *Some normative effects persist over time unless some other and subsequent event terminate them. For example: "If one causes damage, one has to provide compensation.". Other effects hold on the condition and only while the antecedent conditions of the norms hold. For example: "If one is in a public office, one is forbidden to smoke".*

[58] Note that these constructions can give rise to very complex norm dependencies, because we can have that the violation of a single norm can activate other (reparative) norms, which in turn, in case of their violation, refer to other norms, and so forth.

15.1 Further readings

We find it worth concluding by pointing the interested reader to those which we consider the main events and forums in the field, and that can be an excellent source of further information, especially on on-going researches. These are: the biannual DEON[59] and ICAIL[60], the annual JURIX[61], and the Journal of Artificial Intelligence and Law[62].

This work was partially supported by EU H2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No. 690974 for the project *MIREL: MIning and REasoning with Legal texts*.

References

[Alchourrón and Bulygin, 1971] C. E. Alchourrón and E. Bulygin. *Normative Systems*. Springer Verlag, 1971.

[Alchourrón and Bulygin, 1981] C. E. Alchourrón and E. Bulygin. The expressive conception of norms. In Risto Hilpinen, editor, *New Studies in Deontic Logic*, pages 95–125. D. Reidel Publishing Company, Dordrecht, 1981.

[Alchourrón and Bulygin, 1984] Carlos E. Alchourrón and Eugenio Bulygin. Permission and permissive norms. In W. Krawietz et al., editor, *Theorie der Normen*. Duncker & Humblot, 1984.

[Alchourrón and Makinson, 1981] Carlos E. Alchourrón and David C. Makinson. Hierarchies of regulations and their logic. In Risto Hilpinen, editor, *New Studies in Deontic Logic*, pages 125–148. D. Reidel, Dordrecht, 1981.

[Alchourrón and Makinson, 1982] Carlos E. Alchourrón and David C. Makinson. The logic of theory change: Contraction functions and their associated revision functions. *Theoria*, 48:14–37, 1982.

[Alchourrón and Martino, 1990] Carlos E. Alchourrón and Antonio Martino. Logic without truth. *Ratio Juris*, 3:46–67, 1990.

[Alchourrón *et al.*, 1985] Carlos E. Alchourrón, Peter Gärdenfors, and David Makinson. On the logic of theory change: Partial meet contraction and revision functions. *Journal of Symbolic Logic*, 50:510–530, 1985.

[Alchourrón, 1993] C. E. Alchourrón. Philosophical foundations of deontic logic and the logic of defeasible conditionals. In J.-J. Meyer and R. J.Wieringa, editors, *Deontic Logic in Computer Science*. Wiley, New York, 1993.

[Alexy and Dreier, 1991] R. Alexy and R. Dreier. Statutory interpretation in the Federal Republic of Germany. In MacCormick and Summers [1991].

[59] International Conference on Deontic Logic and Normative Systems. See: `http://deonticlogic.org`

[60] International Conference on Artificial intelligence and Law. See the website of the International Association of Artificial Intelligence and Law (IAAIL): `www.iaail.org`

[61] The Foundation for Legal Knowledge Based Systems. Website: `www.jurix.nl`

[62] Website: `www.springer.com/computer/ai/journal/10506`

[Alexy, 1989] R. Alexy. *A Theory of Legal Argumentation*. Clarendon, Oxford, 1989.

[Alexy, 2002] Robert Alexy. *A Theory of Constitutional Rights*. Oxford University Press, Oxford, 2002.

[Antoniou *et al.*, 2001] Grigoris Antoniou, David Billington, Guido Governatori, and Michael J. Maher. Representation results for defeasible logic. *ACM Transactions on Computational Logic*, 2:255–287, 2001.

[Bench-Capon *et al.*, 2000] Trevor J. M. Bench-Capon, T. Geldard, and Paul H. Leng. A method for the computational modelling of dialectical argument with dialogue games. *Artif. Intell. Law*, 8(2/3):233–254, 2000.

[Billington *et al.*, 2010] David Billington, Grigoris Antoniou, Guido Governatori, and Michael J. Maher. An inclusion theorem for defeasible logic. *ACM Transactions in Computational Logic*, 12(1):article 6, 2010.

[Bobbio, 1958] N. Bobbio. *Teoria della norma giuridica*. Giappichelli, 1958.

[Boella and van der Torre, 2003a] G. Boella and L. van der Torre. Permissions and obligations in hierarchical normative systems. In *ICAIL '03*, pages 109–118. ACM, 2003.

[Boella and van der Torre, 2003b] G. Boella and L. van der Torre. Permissions and undercutters. In *NRAC'03*, pages 51–57, Acapulco, 2003.

[Boella and Van der Torre, 2004] G. Boella and L. Van der Torre. Regulative and constitutive norms in normative multiagent systems. In D. Dubois, C. A. Christopher A. Welty, and M. Williams, editors, *Proceedings of KR2004, Whistler, Canada*, pages 255–266, 2004.

[Boella and van der Torre, 2005] G. Boella and L. van der Torre. Permission and authorization in normative multiagent systems. In *ICAIL '05*, pages 236–237. ACM, 2005.

[Boella *et al.*, 2009] Guido Boella, Gabriella Pigozzi, and Leendert van der Torre. A normative framework for norm change. In *8th International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS 2009)*, pages 169–176. IFAAMAS, 2009.

[Boella *et al.*, 2010] Guido Boella, Guido Governatori, Antonino Rotolo, and Leendert van der Torre. Lex minus dixit quam voluit, lex magis dixit quam voluit: A formal study on legal compliance and interpretation. In *Proceedings of AICOL*. Springer, 2010.

[Brown, 2000] M.A. Brown. Conditional obligation and positive permission for agents in time. *Nordic Journal of Philosophical Logic*, 5(2):83–111, 2000.

[Brozek, 2013] Bartosz Brozek. Legal interpretation and coherence. In Michal Araszkiewicz and Jaromir Savelka, editors, *Coherence: Insights from Philosophy, Jurisprudence and Artificial Intelligence*. Springer, 2013.

[Calardo *et al.*, 2018] Erica Calardo, Guido Governatori, and Antonino Rotolo. Sequence semantics for modelling reason-based preferences. *Fundamenta Informaticae*, 158:217–238, 2018.

[Cristani *et al.*, 2017] Matteo Cristani, Francesco Olivieri, and Antonino Rotolo. Changes to temporary norms. In *Proceedings of the 16th edition of the*

International Conference on Articial Intelligence and Law, ICAIL 2017, London, United Kingdom, June 12-16, 2017, pages 39–48, 2017.

[Dung, 1995] Phan Minh Dung. On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. *Artificial Intelligence*, 77(2):321–358, 1995.

[Elgesem, 1997] Dag Elgesem. The modal logic of agency. *Nordic Journal of Philosophical Logic*, 2:1–48, 1997.

[Gabbay *et al.*, 2003] Dov M. Gabbay, Gabriella Pigozzi, and John Woods. Controlled revision - an algorithmic approach for belief revision. *Journal of Logic and Computation*, 13(1):3–22, 2003.

[Gelati *et al.*, 2004] Jonathan Gelati, Antonino Rotolo, Giovanni Sartor, and Guido Governatori. Normative autonomy and normative co-ordination: Declarative power, representation, and mandate. *Artif. Intell. Law*, 12(1-2):53–81, 2004.

[Goldman, 1970] A.I. Goldman. *A theory of human action*. Prentice-Hall, 1970.

[Gordon *et al.*, 2007] Thomas F. Gordon, Henry Prakken, and Douglas Walton. The Carneades model of argument and burden of proof. *Artificial Intelligence*, 171(10-11):875–896, 2007.

[Gordon, 1995] Thomas F. Gordon. *The Pleadings Game: An Artificial Intelligence Model of Procedural Justice*. Kluwer, Dordrecht, 1995.

[Governatori and Rotolo, 2005] Guido Governatori and Antonino Rotolo. On the axiomatization of Elgesem's logic of agency and ability. *Journal of Philosophical Logic*, 34(4):403–431, 2005.

[Governatori and Rotolo, 2006] G. Governatori and A. Rotolo. Logic of violations: A Gentzen system for reasoning with contrary-to-duty obligations. *Australasian Journal of Logic*, 4:193–215, 2006.

[Governatori and Rotolo, 2008a] Guido Governatori and Antonino Rotolo. Bio logical agents: Norms, beliefs, intentions in defeasible logic. *Journal of Autonomous Agents and Multi Agent Systems*, 17(1):36–69, 2008.

[Governatori and Rotolo, 2008b] Guido Governatori and Antonino Rotolo. A computational framework for institutional agency. *Artificial Intelligence and Law*, 16(1):25–52, 2008.

[Governatori and Rotolo, 2010] Guido Governatori and Antonino Rotolo. Changing legal systems: Legal abrogations and annulments in defeasible logic. *The Logic Journal of IGPL*, 18(1):157–194, 2010.

[Governatori and Rotolo, 2013] Guido Governatori and Antonino Rotolo. Computing temporal defeasible logic. In Leora Morgenstern, Petros S. Stefaneas, François Lévy, Adam Wyner, and Adrian Paschke, editors, *7th International Symposium on Theory, Practice, and Applications of Rules on the Web (RuleML 2013)*, volume 8035 of *LNCS*, pages 114–128, 2013.

[Governatori and Sadiq, 2008] G. Governatori and S. Sadiq. The journey to business process compliance. *Handbook of Research on BPM*, pages 426–454, 2008.

[Governatori and Sartor, 2010] Guido Governatori and Giovanni Sartor. Bur-

dens of proof in monological argumentation. In Radboud Winkels, editor, *The Twenty-Third Annual Conference on Legal Knowledge and Information Systems (Jurix 2010)*, pages 57–66, Amsterdam, 2010. IOS Press.

[Governatori and Shek, 2012] G. Governatori and S. Shek. Rule based business process compliance. In *Proceedings of the RuleML2012@ECAI Challenge*, number 874 in CEUR, page 5, 2012.

[Governatori *et al.*, 2004] Guido Governatori, Michael J. Maher, David Billington, and Grigoris Antoniou. Argumentation semantics for defeasible logics. *Journal of Logic and Computation*, 14(5):675–702, 2004.

[Governatori *et al.*, 2005a] Guido Governatori, Monica Palmirani, Régis Riveret, Antonino Rotolo, and Giovanni Sartor. Norm modifications in defeasible logic. In Marie-Francine Moens and Peter Spyns, editors, *The Eighteenth Annual Conference on Legal Knowledge and Information Systems (Jurix 2005)*, pages 13–22, Amsterdam, 2005. IOS Press.

[Governatori *et al.*, 2005b] Guido Governatori, Antonino Rotolo, and Giovanni Sartor. Temporalised normative positions in defeasible logic. In Anne Gardner, editor, *10th International Conference on Artificial Intelligence and Law (ICAIL 2005)*, pages 25–34. ACM Press, 2005.

[Governatori *et al.*, 2007a] Guido Governatori, Joris Hulstijn, Régis Riveret, and Antonino Rotolo. Characterising deadlines in temporal modal defeasible logic. In Mehmet A. Orgun and John Thornton, editors, *20th Australian Joint Conference on Artificial Intelligence*, volume 4830 of *Lecture Notes in Artificial Intelligence*, pages 486–496, Heidelberg, 2007. Springer.

[Governatori *et al.*, 2007b] Guido Governatori, Antonino Rotolo, Régis Riveret, Monica Palmirani, and Giovanni Sartor. Variants of temporal defeasible logic for modelling norm modifications. In Radboud Winkels, editor, *Proceedings of 11th International Conference on Artificial Intelligence and Law (ICAIL 2007)*, pages 155–159, New York, 2007. ACM Press.

[Governatori *et al.*, 2012] G. Governatori, A. Rotolo, and E. Calardo. Possible world semantics for defeasible deontic logic. In T. Ågotnes, J. Broersen, and D. Elgesem, editors, *DEON 2012*, volume 7393 of *LNCS*, pages 46–60. Springer, 2012.

[Governatori *et al.*, 2013a] Guido Governatori, Francesco Olivieri, Antonino Rotolo, and Simone Scannapieco. Computing strong and weak permissions in defeasible logic. *Journal of Philosophical Logic*, 42(6):799–829, 2013.

[Governatori *et al.*, 2013b] Guido Governatori, Antonino Rotolo, Francesco Olivieri, and Simone Scannapieco. Legal contractions: a logical analysis. In *ICAIL*, pages 63–72, 2013.

[Governatori *et al.*, 2016a] Guido Governatori, Francesco Olivieri, Erica Calardo, and Antonino Rotolo. Sequence semantics for norms and obligations. In *Proceedings of DEON 2016*, London, 2016. College Publications.

[Governatori *et al.*, 2016b] Guido Governatori, Francesco Olivieri, Erica Calardo, Antonino Rotolo, and Matteo Cristani. Sequence semantics for normative agents. In *PRIMA 2016: Principles and Practice of Multi-Agent*

Systems - 19th International Conference, Phuket, Thailand, August 22-26, 2016, Proceedings, pages 230–246, 2016.

[Governatori, 2005] Guido Governatori. Representing business contracts in RuleML. *International Journal of Cooperative Information Systems*, 14(2-3):181–216, 2005.

[Governatori, 2011] Guido Governatori. On the relationship between Carneades and Defeasible Logic. In Kevin Ashley, editor, *The 13th International Conference on Artificial Intelligence and Law (ICAIL 2011)*, pages 31–40. ACM, 2011.

[Governatori, 2015] Guido Governatori. Thou Shalt is not You Will. In Katie Atkinson, editor, *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Law (ICAIL 2015)*, pages 63–68. ACM, 2015.

[Greenberg, 2004] Mark Greenberg. How facts make law. *Legal Theory*, 10(3):157–198, 2004.

[Grossi and Jones, 2013a] D. Grossi and A. Jones. Constitutive rules and counts-as conditionals. In X. Parent D. Gabbay, J. Horty and L. van der Torre, editors, *Deontic Logic Handbook*. College Publications, London, 2013.

[Grossi and Jones, 2013b] D. Grossi and A. J. I. Jones. Constitutive norms and counts-as conditionals. Handbook of Deontic Logic, 2013.

[Grossi and Rotolo, 2011] Davide Grossi and Antonino Rotolo. Logic in law: A concise overview. In Dov M. Gabbay, editor, *Logic and philosophy today. - Vol. 2*, pages 251–274. College Publications, London, 2011.

[Guastini, 1998] Riccardo Guastini. *Teoria e dogmatica delle fonti*. Giuffré, Milan, 1998.

[Hage, 1997] J. Hage. *Reasoning with Rules: An Essay on Legal Reasoning and Its Underlying Logic*. Kluwer, Dordrecht, 1997.

[Hansen *et al.*, 2007] J. Hansen, G. Pigozzi, and L. van der Torre. Ten philosophical problems in deontic logic. In Guido Boella, Leon van der Torre, and Harko Verhagen, editors, *Normative Multi-agent Systems*, number 07122 in DROPS Proceedings. Internationales Begegnungs- und Forschungszentrum fuer Informatik (IBFI), Schloss Dagstuhl, Germany, 2007.

[Hansen, 2005] Jörg Hansen. Conflicting imperatives and dyadic deontic logic. *J. Applied Logic*, 3(3-4):484–511, 2005.

[Hansson, 1969] B. Hansson. An analysis of some deontic logics. *Nous*, (3):373–398, 1969.

[Hart, 1951] H.L.A. Hart. The ascription of responsibility and rights. In A. Flew, editor, *Logic and Language*. Blackwell, 1951.

[Hart, 1994] H.L.A. Hart. *The Concept of Law*. Clarendon, Oxford, 1994.

[Herrestad and Krogh, 1995] Henning Herrestad and Christen Krogh. Obligations directed from bearers to counterparts. In *Proceedings of the 5th International Conference on Artificial Intelligence and Law*, ICAIL '95, pages 210–218, New York, NY, USA, 1995. ACM.

[Hohfeld, 1913] Wesley Necomb Hohfeld. Some fundamental legal conceptions

as applied in judicial reasoning. In *Yale Law Journal*, pages 16–59, 1913.

[Hohfeld, 1917] Wesley Necomb Hohfeld. Fundamental legal conceptions as applied in judicial reasoning. *Yale Law Journal*, 26:710–770, 1917.

[Horovitz, 1972] J. Horovitz. *Law and Logic*. Springer, 1972.

[Horty, 2002] John F. Horty. Skepticism and floating conclusions. *Artif. Intell.*, 135:55–72, February 2002.

[Jones and Sergot, 1996] Andrew J. I. Jones and Marek J. Sergot. A formal characterisation of institutionalised power. *Logic Journal of the IGPL*, 4(3):427–443, 1996.

[Jørgensen, 1937 1938] J. Jørgensen. Imperatives and logic. *Erkenntnis*, 4:288–296, 1937–1938.

[Kalinowski, 1953] G. Kalinowski. Théorie des propositions normatives. *Studia logica*, 17:147–82, 1953.

[Kalinowski, 1959] G. Kalinowski. Y a-t-il une logique juridique? *Logique et Analyse*, 5:48–53, 1959.

[Kalinowski, 1972] G. Kalinowski. *La logique des normes*. PUF, Paris, 1972.

[Kanger and Kanger, 1966] S. Kanger and H. Kanger. Rights and parliamentarism. *Theoria*, 32(2):85–129, 1966.

[Kanger, 1957] S. Kanger. A note on quantification and modalities. *Theoria*, 23:133–134, 1957.

[Kelsen, 1967] H. Kelsen. *Pure Theory of Law*. California library reprint series. University of California Press, 1967.

[Kelsen, 1979] Hans Kelsen. *Allgemeine Theorie der Normen*. Manz, Vienna, 1979.

[Kelsen, 1991] Hans Kelsen. *General Theory of Norms*. Clarendon, Oxford, 1991.

[Lindahl and Odelstad, 2000] L. Lindahl and J. Odelstad. An algebraic analysis of normative systems. *Ratio Juris*, 13:261–278, 2000.

[Lindahl and Odelstad, 2008] L. Lindahl and J. Odelstad. Intermediaries and intervenients in normative systems. *Journal of Applied Logic*, 6(2):229–258, 2008.

[Lindahl, 1977] Lars Lindahl. *Position of change: A Study in law and logic*. Reidel, Dordrecht, 1977.

[Lodder, 1999] Arno R. Lodder. *DiaLaw: On Legal Justification and Dialogical Models of Argumentation*. Kluwer, Dordrecht, 1999.

[Macagno *et al.*, 2012] F. Macagno, G. Sartor, and D. Walton. Argumentation schemes for statutory interpretation. In *Proc. ARGUMENTATION 2012*. Masaryk University, 2012.

[MacCormick and Summers, 1991] D.N. MacCormick and R.S. Summers, editors. *Interpreting Statutes: A Comparative Study*. Ashgate, 1991.

[MacCormick, 2005] Neil MacCormick. *Rhetoric and the Rule of Law: A Theory of Legal Reasoning*. Oxford University Press, Oxford, 2005.

[Makinson and van der Torre, 2000] D. Makinson and L. van der Torre. Input-

output logics. *Journal of Philosophical Logic*, 29(4):383–408, 2000.

[Makinson and van der Torre, 2001] David Makinson and Leendert van der Torre. Constraints for input/output logics. *J. Philosophical Logic*, 30(2):155–185, 2001.

[Makinson and van der Torre, 2003] D. Makinson and L. van der Torre. Permission from an input/output perspective. *Journal of Philosophical Logic*, 32(4):391–416, 2003.

[Makinson, 1986] David Makinson. On the formal representation of rights relations. *Journal of Philosophical Logic*, 15:403–25, 1986.

[Makinson, 1999a] D. Makinson. On a fundamental problem of deontic logic. In P. McNamara and H. Prakken, editors, *Norms, Logics and Information Systems. New Studies in Deontic Logic and Computer Science*, volume 49 of *Frontiers in Artificial Intelligence and Applications*, pages 29–53. IOS Press, Amsterdam, 1999.

[Makinson, 1999b] David Makinson. On a fundamental problem of deontic logic. In P. McNamara and H. Prakken, editors, *Norms, Logics, and Information Systems*. IOS Press, 1999.

[Makinson, 2005] David Makinson. *Bridges from Classical to Nonmonotonic Logic*. King's College Publications, London, 2005.

[Malerba *et al.*, 2016] Alessandra Malerba, Antonino Rotolo, and Guido Governatori. Interpretation across legal systems. In *Legal Knowledge and Information Systems - JURIX 2016: The Twenty-Ninth Annual Conference*, pages 83–92. IOS Press, 2016.

[Nino, 2013] Carlos Santiago Nino. *Introducción al análisis del Derecho*. Ariel, Buenos Aires, 2013.

[Nute, 1994] Donald Nute. Defeasible logic. In Dov M Gabbay, Christopher John Hogger, and John Alan Robinson, editors, *Handbook of Logic in Artificial Intelligence and Logic Programming*, volume 3, pages 353–395. Oxford University Press, Oxford, 1994.

[Nute, 1998] D. Nute. Norms, priorities, and defeasibility. In H. Prakken and P. McNamara, editors, *Norms, Logics and Information Systems*. IOS Press, 1998.

[Opalek and Wolenski, 1991] K. Opalek and J. Wolenski. Normative systems, permission and deontic logic. *Ratio Juris*, 4:334–348, 1991.

[Pattaro, 2005] E. Pattaro. *The Law and The Right: A Reappraisal of the Reality that Ought to Be*. Springer, Dordrecht, 2005.

[Peczenik, 1989] A. Peczenik. *On law and reason*. Kluwer, Dordrecht, 1989.

[Pollock, 1995] John L. Pollock. *Cognitive Carpentry: A Blueprint for how to Build a Person*. MIT Press, Cambridge, MA, USA, 1995.

[Pörn, 1977] Ingmar Pörn. *Action Theory and Social Science: Some Formal Models*. Reidel, Dordrecht, 1977.

[Prakken and Sartor, 1996] Henry Prakken and Giovanni Sartor. A dialectical model of assessing conflicting arguments in legal reasoning. *Artif. Intell. Law*,

4(3-4):331–368, 1996.

[Prakken and Sartor, 1997] Henry Prakken and Giovanni Sartor. Argument-based extended logic programming with defeasible priorities. *Journal of Applied Non-Classical Logics*, 7(1), 1997.

[Prakken and Sartor, 2002] Henry Prakken and Giovanni Sartor. The role of logic in computational models of legal argument: A critical survey. In *Computational Logic: Logic Programming and Beyond*, pages 342–381. Springer, 2002.

[Prakken and Sartor, 2004] Henry Prakken and Giovanni Sartor. The three faces of defeasibility in the law. *Ratio Juris*, 17:118–139, 2004.

[Prakken and Sartor, 2009] H. Prakken and G. Sartor. A logical analysis of burdens of proof. In H. Kaptein, editor, *Legal Evidence and Proof: Statistics, Stories, Logic*. Ashgate, 2009.

[Prakken and Sartor, 2013] H. Prakken and G. Sartor. Formalising arguments about norms. In *Proc. JURIX 2013*. IOS Press, 2013.

[Prakken and Sartor, 2015] Henry Prakken and Giovanni Sartor. Law and logic: A review from an argumentation perspective. *Artif. Intell.*, 227:214–245, 2015.

[Prakken and Sergot, 1996] Henry Prakken and Marek J. Sergot. Contrary-to-duty obligations. *Studia Logica*, 57(1):91–115, 1996.

[Prakken and Vreeswijk, 2002] H. Prakken and G.A.W. Vreeswijk. Logics for defeasible argumentation. In D.M. Gabbay and F. Guenthner, editors, *Handbook of Philosophical Logic, 2nd edition*, volume 4, pages 219–318. Kluwer, 2002.

[Prakken, 1995] H. Prakken. *Logical Tools for Modelling Legal Argument: A Study of Defeasible Reasoning in Law*. Kluwer, Dordrecht, 1995.

[Prakken, 1996] Henry Prakken. Two approaches to the formalisation of defeasible deontic reasoning. *Studia Logica*, 57(1):73–90, 1996.

[Prakken, 2001] H Prakken. Relating protocols for dynamic dispute with logics for defeasible argumentation. *Synthese*, 127(1999):187–219, 2001.

[Rawls, 1955] John Rawls. Two concepts of rules. *The Philosophical Review*, 64(1):3–32, 1955.

[Rawls, 1971] J. Rawls. *A Theory of Justice*. Harvard University Press, Cambridge, MA, 1971.

[Raz, 1990] Joseph Raz. *Practical Reason and Norms*. Princeton University Press, Princeton, 1990.

[Rescher, 1977] N. Rescher. *Dialectics: A Controversy-Oriented Approach to the Theory of Knowledge*. State University of New York Press, New York, 1977.

[Ross, 1958] A. Ross. *On Law and Justice*. Stevens, London, 1958.

[Ross, 1968] Alf Ross. *Directives and Norms*. Routledge, London, 1968.

[Rotolo *et al.*, 2015] Antonino Rotolo, Guido Governatori, and Giovanni Sartor. Deontic defeasible reasoning in legal interpretation: two options for modelling

interpretive arguments. In *Proceedings of the 15th International Conference on Artificial Intelligence and Law, ICAIL 2015, San Diego, CA, USA, June 8-12, 2015*, pages 99–108, 2015.

[Rotolo, 2010] Antonino Rotolo. Retroactive legal changes and revision theory in defeasible logic. In Guido Governatori and Giovanni Sartor, editors, *Proceedings of the 10th International Conference on Deontic Logic in Computer Science (DEON 2010)*, volume 6181 of *LNAI*, pages 116–131. Springer, 2010.

[Royakkers and Dignum, 1997] L. M. M. Royakkers and F. Dignum. The logic of enactment. In *ICAIL*, 1997.

[Rubino *et al.*, 2006] Rossella Rubino, Antonino Rotolo, and Giovanni Sartor. An OWL ontology of fundamental legal concepts. In *Legal Knowledge and Information Systems - JURIX 2006: The Nineteenth Annual Conference on Legal Knowledge and Information Systems, Paris, France, 7-9 December 2006*, pages 101–110, 2006.

[Ruiter, 2001] D.W. Ruiter. *Legal Institutions*. Law and Philosophy Library. Springer, Dordrecht, 2001.

[Santos *et al.*, 1997] Filipe Santos, Andrew J.I. Jones, and José Carmo. Action concepts for describing organised interaction. In *Thirtieth Annual Hawaii International Conference on System Sciences*. IEEE Computer Society Press, Los Alamitos, 1997.

[Sartor *et al.*, 2014] Giovanni Sartor, Doug Walton, Fabrizio Macagno, and Antonino Rotolo. Argumentation schemes for statutory interpretation: A logical analysis. In *Proc. JURIX 2014*. IOS Press, 2014.

[Sartor, 1995] G. Sartor. Defeasibility in legal reasoning. In Z. Bankowski, I. White, and U. Hahn, editors, *Informatics and the foundations of legal reasoning*. Kluwer, 1995.

[Sartor, 2005] G. Sartor. *Legal Reasoning: A Cognitive Approach to the Law*. Springer, 2005.

[Sartor, 2010] Giovanni Sartor. Doing justice to rights and values: teleological reasoning and proportionality. *Artif. Intell. Law*, 18(2):175–215, June 2010.

[Searle, 1969] J. R. Searle. *Speech Acts: An Essay in the Philosophy of Language*. Cambridge University Press, Cambridge, 1969.

[Searle, 1995] J. R. Searle. *The Construction of Social Reality*. Penguin, Harmondsworth, 1995.

[Sergot, 2001] Marek J. Sergot. A computational theory of normative positions. *ACM Trans. Comput. Log.*, 2(4):581–622, 2001.

[Stenius, 1963] Erik Stenius. Principles of a logic of normative systems. *Acta Philosophica Fennica*, 16:247–260, 1963.

[Stolpe, 2010a] A. Stolpe. Relevance, derogation and permission. In *DEON*, pages 98–115. Springer, 2010.

[Stolpe, 2010b] A. Stolpe. A theory of permission based on the notion of derogation. *J. Applied Logic*, 8(1):97–113, 2010.

[Stolpe, 2010c] Audun Stolpe. Norm-system revision: theory and application.

Artificial Intelligence and Law, 18(3):247–283, 2010.

[Tarello, 1980] G. Tarello. *L'interpretazione della legge*. Giuffrè, 1980.

[van Benthem *et al.*, 2013] Johan van Benthem, Davide Grossi, and Fenrong Liu. Priority structures in deontic logic. *Theoria*, 2013.

[van der Torre, 1997] Leendert van der Torre. *Reasoning about obligations: defeasibility in preference-based deontic logic*. PhD thesis, Erasmus University Rotterdam, 1997.

[von Wright, 1951] G. H. von Wright. Deontic logic. *Mind*, 60(237):1–15, 1951.

[von Wright, 1963] G. H. von Wright. *Norm and Action: A Logical Inquiry*. Routledge, London, 1963.

[Walton and Krabbe, 1995] D.N. Walton and E.C.W. Krabbe. *Commitment in Dialogue: Basic Concepts of Interpersonal Reasoning*. SUNY series in logic and language. State University of New York Press, 1995.

[Weinberger, 1981] Ota Weinberger. *Normentheorie als Grundlage der Jurisprudenz und der Ethik. Eine Auseinandersetzung mit Hans Kelsens Theorie der Normen*. Duncker & Humblot, Berlin, 1981.

[Weinberger, 1989] Ota Weinberger. *Rechtslogik*. Duncker & Humblot, Berlin, 1989.

[Wheeler and Alberti, 2011] Gregory R. Wheeler and Marco Alberti. No revision and no contraction. *Minds and Machines*, 21(3):411–430, 2011.

Guido Governatori
Data61, CSIRO, Australia
Email: guido.governatori@data61.csiro.au

Antonino Rotolo
CIRSFID and Department of Legal Studies, University of Bologna, Italy
Email: antonino.rotolo@unibo.it

Giovanni Sartor
CIRSFID and Department of Legal Studies, University of Bologna, Italy
Email: giovanni.sartor@unibo.it

10
Deontic Logic and Game Theory

Olivier Roy

Abstract. This chapter provides an overview of applications of deontic logic for analyzing norms of rationality in strategic interaction, i.e. the subject-matter of game theory. The main take-home message is that even within the realm of classical, non-cooperative game theory, there are many modeling options for representing the rational obligations and permissions of the players. Different choices will result in different logical behaviors for obligation and permission, each of which comes with certain benefits and drawbacks. The goal of this entry is to highlight some of these modeling choices. It starts with a very brief introduction to game theory and then moves on to its application in deontic logic.

1 Games in strategic forms **762**

2 Normative interpretation of solution concepts **764**

3 Obligations and permissions in games **765**

4 Alternative 1: Standard deontic logic **766**

5 Alternative 2: Optimal and best **769**

6 Alternative 3: Permissions as sufficient conditions **771**

7 Conclusion **775**

The work on this chapter has been partly supported by the DFG-NCN "PIOTR Project" (RO 4548/4-1). The author would also like to thank the two anonymous reviewers, John F. Horty, Dominik Klein, Alessandra Marra, Frederik Van De Putte, as well as the participants to the *Trends in Logic XVII 2017* conference in Lubin and the 14^{th} *Deontic Logic and Normative System (DEON 2018)* in Utrecht for useful comments and suggestions.

Handbook of Deontic Logic and Normative Systems, Volume 2

1 Games in strategic forms

This chapter considers only one-shot games where the players choose simultaneously. In game-theoretic jargon they are called games in "strategic" or "normal forms". They can also be seen as representations that abstract away from any sequential or temporal structure the "real" game might have. For a general introduction to game theory, see for instance [Osborne and Rubinstein, 1994].

Here is a simple example that will follow us throughout:

Ann\Bob	L	R
U	1,1	0,0
D	0,1	1,0

In this game there are two players, Ann and Bob. Ann can choose the upper (U) or the lower (D) row, and Bob either the left (L) or the right (R) column. Each combination of these choices, for instance, (U, L) or (D, R), is called a "strategy profile" and corresponds to one possible outcome of the game. Ann and Bob have preferences over these outcomes, which are expressed by the numbers in each cell of the matrix, with 1 preferred to 0. The left-hand numbers represent Ann's preferences, the right-hand ones Bob's. Throughout this chapter I will follow the mainstream decision- and game-theoretic interpretation of these numbers, namely in terms of interval-scale utility functions representing the agents' preferences. These, in turn, are typically interpreted behaviorally. We read Ann's preferring outcome x to y as saying that she would choose x if she was given a choice between the two. Under that interpretation, rationality, and by the same token the normative interpretation of solution concepts (Section 2), becomes a matter of choosing coherently.

The general definition of a game in strategic form goes as follows:

Definition 1.1. *A game in strategic form $\mathcal{G}$ is a tuple $\langle \mathcal{A}, (S_i, \pi_i)_{i \in \mathcal{A}} \rangle$ where*

- *$\mathcal{A}$ is a finite set of agents.*
- *For each agent $i \in \mathcal{A}$, S_i is a finite set of strategies for i.*
  - *An element σ of the Cartesian product $\prod_{i \in \mathcal{A}} S_i$ of the strategy sets for all players is called a "strategy profile".*
  - *σ_i is i's strategy in σ.*
  - *σ_{-i} is the profile σ for all agents except i.*

- $\pi_i \colon \prod_{j \in \mathcal{A}} S_j \to \mathbb{R}$ *is a payoff function for agent* i.

For most of the discussion below we will only use the qualitative preference orders $\succeq_i$ induced by the π_i's in the natural way:

$$\sigma \succeq_i \sigma' \;\Leftrightarrow\; \pi_i(\sigma) \geq \pi_i(\sigma').$$

The strict version of this relation is defined by the condition that $\sigma \succ_i \sigma'$ if and only if $\sigma \succeq_i \sigma'$ but not the other way around.

Rational plays are defined using so-called solution concepts. Let us go back to our example. What Ann's best response is depends on what Bob does. If he plays L then her best option is to play U. If, however, he plays R then her best response is D. Things are different for Bob. What is the best response for him does *not* depend on what Ann does. If she plays U he gets 1 by playing L, and 0 by playing R; and the same holds if Ann plays D. In game-theoretic terminology we say that R is "strictly dominated by L". The solution concept of (iterated) elimination of strictly dominated strategies then starts by removing R from the set of strategies to consider, leaving L as the only option for Bob, and then eliminates every strategy that is dominated in the reduced game, if any. Here, indeed, there is one. As we have already seen, if we consider L as the only possible action for Bob, the best response for Ann is to play U, leaving (U, L) as the only solution of this game. This profile also happens to be a so-called Nash equilibrium. Neither Ann nor Bob have an incentive to unilaterally deviate from it. We have already seen that Bob has no incentive whatsoever to play R instead of L. For Ann, given that Bob plays L, switching to D would yield a strictly lower payoff for her. In fact, (U, L) is a so-called strict equilibrium, that is, Ann and Bob both have an incentive to unilaterally conform to it.

More generally, a solution concept $\mathbb{S}$ assigns to each game $\mathcal{G}$ a set of strategy profiles $\mathbb{S}(\mathcal{G}) \subseteq \prod_i S_i$. For illustrative purposes I will here only consider Nash equilibrium and iterated elimination of strictly dominated strategies (IESDS). The first is probably the most widely used solution concept, while the second has the clearer decision-theoretic foundation [Pacuit and Roy, 2017]. There is of course a plethora of other solution concepts (for details, see again [Osborne and Rubinstein, 1994]), but I will leave them aside here.

Definition 1.2. *Let $\mathcal{G}$ be a game in strategic form. Then the strategy s_i of agent i is strictly dominated whenever there is another strategy s_i' of i such that for all profiles σ_{-i},*

$$(s_i', \sigma_{-i}) \succ_i (s_i, \sigma_{-i}).$$

The set SD^ω of strategy profiles is defined inductively as follows:

- $SD^0 = \prod_i S_i$,
- $SD^{n+1} = SD^n / \bigcup_{i \in \mathcal{A}} \{s_i \colon s_i$ *is strictly dominated in the restriction of* $\mathcal{G}$ *to* $SD^n\}$,
- $SD^\omega = \bigcap_{n<\omega} SD^n$.

The set SD^ω, i.e., the set of profiles that survive iterated elimination of strictly dominated strategies, is called the "IESDS set".

Definition 1.3. *Let $\mathcal{G}$ be a game in strategic form. Then σ is a Nash equilibrium iff for all i and $s' \in S_i$,*

$$\sigma \succeq_i (s', \sigma_{-i}).$$

2 Normative interpretation of solution concepts

In this chapter I will be focusing on the so-called *normative* interpretation of solution concepts like IESDS or Nash equilibrium [De Bruin, 2010; Pacuit and Roy, 2017]. Under that interpretation, IESDS specifies the *rational* strategies for each agent to play. I take strategic rationality to be one source of recommendations in games, but I remain noncommittal regarding the relationship between it and other potential sources of recommendations, for instance, morality or prudence. In particular, by using games and solution concepts to interpret deontic operators, I am not suggesting a "moral interpretation" of game-theoretical solution concepts, as suggested, for instance, in [Tamminga, 2013].

The normative interpretation of the Nash equilibrium solution concept is different, however, because being a Nash equilibrium is a property of a strategy *profile*. Since in non-trivial cases players cannot unilaterally achieve particular profiles, it doesn't make sense to recommend to each player, individually, to play a Nash equilibrium. Another way to see this is that non-equilibrium outcomes can occur even though all players play a strategy that *might* lead to an equilibrium. See [Tamminga and Duijf, 2017] for the importance of this fact in the study of collective obligations. Pure coordination games provide a very simple example of that:

Ann\Bob	A	B
a	1,1	0,0
b	0,0	1,1

Here, for each player, both actions are compatible with an equilibrium, but of course if Ann plays a and Bob plays B they both have an incentive to unilaterally deviate. For that reason, a more plausible normative interpretation of Nash equilibrium is that this solution concept singles out the profiles that are expected, in an intuitive sense, to result from the interaction of rational players.

More abstractly, given a particular solution concept $\mathbb{S}$ we take the set $\mathbb{S}(\mathcal{G})$ to provide an *extensional definition* of the rational profiles or strategies in $\mathcal{G}$. For Nash equilibrium, $\mathbb{S}(\mathcal{G})$ is thus the set of equilibrium profiles, and for IESDS we have $\mathbb{S}(\mathcal{G}) = SD^{\omega}$.

Taking the set $\mathbb{S}(\mathcal{G})$ to be an extensional definition of the set of rational strategy profiles means that for a particular profile σ to be rational it is both necessary and sufficient that it is a member of $\mathbb{S}(\mathcal{G})$. No profile outside of $\mathbb{S}(\mathcal{G})$ is a rational one, and nothing else than being a member of that set is required for a profile to be rational. For solution concepts like IESDS this can be transferred to each player's strategies. If, for instance, the profile σ is in SD^{ω} we can say that playing σ_i is rational for i, and that any strategy s_i' that is not part of a profile in SD^{ω} is not rational for i.

3 Obligations and permissions in games

Even starting from the given solution concept $\mathbb{S}$, there are many modeling options for interpreting a deontic language. The language of standard deontic logic that I will use throughout is an extension of classical propositional logic with two modal operators O and P, expressing respectively that a certain proposition is obligatory or permitted. One could of course extend this language with agent-, groups- or coalitions-relative obligations, or even with different solution concepts, by using a family of modal operators instead. A formula A has the following form:

$$A := p \mid \neg A \mid A \wedge A \mid \mathrm{O}A \mid \mathrm{P}A.$$

In standard deontic logic, O and P are usually duals, that is, $\mathrm{O}A \leftrightarrow \neg\mathrm{P}\neg A$, but this duality fails for some of the interpretations of O and P that we will be studying below. So both operators are taken as primitives here. Later on, this language will be extended with additional technical tools, for instance, the so-called universal modality $\Box$, which expresses the fact that something is necessary or settled in a particular situation.

This language is usually interpreted using deontic models of the form

$$\mathcal{M} = \langle W, R, v \rangle$$

where R is a binary relation connecting each state $w \in W$ to all the states w' that are ideal from the perspective of w, and v is a valuation function assigning to each atomic proposition the set of states where it is true. To provide a game-theoretic interpretation of our deontic language we will have to decide how to turn games in strategic form into deontic models, as well as provide truth conditions for the deontic operators in these models. In other words, studying the logical structure of rational recommendations stemming from solution concepts in games can be seen as answering the following two questions:

1. How does one construct a model $\mathcal{M}$ from a given $\mathcal{G}$ and $\mathbb{S}$?
2. What is the semantics of the deontic modalities O and P in such models?

One of the key take-home messages of this chapter is that there are a number of different ways to answer these questions, each yielding a different deontic logic of rational recommendations.

4 Alternative 1: Standard deontic logic

A simple way to start is to use the set of rational solutions to construct a uniform deontic accessibility relation R, i.e., one in which the set of ideal worlds is the same at each w, and take the standard interpretation of O and P. More precisely:

Definition 4.1. *Let $\mathcal{G}$ be a game in strategic form and $\mathbb{S}(\mathcal{G})$ the set of rational strategy profiles for that game. The frame $\mathcal{F}_{\mathcal{G}} = \langle W, R \rangle$ is constructed as follows:*

- *W is a set of states.*
- *$\sigma \colon W \to \prod_i S_i$ is a function assigning a strategy profile to each state in W.*
- *For all w, $w' \in W$, wRw' iff $\sigma(w') \in \mathbb{S}(\mathcal{G})$.*

A model $\mathcal{M}_{\mathcal{G}}$ is a frame $\mathcal{F}_{\mathcal{G}}$ augmented with a valuation v assigning to each proposition p a subset of W.

In this construction each state in W gets assigned a strategy profile in the underlying game $\mathcal{G}$. The intuitive idea is that this $\sigma(w)$ is the profile that is played at state w. The function σ needs not to be surjective nor injective. Some profiles might not be played in any state of the model,

and while others might be played at more than one state. The definition of the relation R then fixes the set of ideal states uniformly across the model: at all states the ideal states are those where a rational profile is played, with "rational" being defined by the solution concept at hand.

We can then use the same semantics for O as given above, and give PA its dual interpretation, i.e., as a shorthand for $\neg O \neg A$.

Definition 4.2. *Let $\mathcal{M}_\mathcal{G}$ be as above. Then let*

$$\mathcal{M}_\mathcal{G}, w \models OA \quad \textit{iff} \quad \mathcal{M}_\mathcal{G}, w' \models A \ \textit{ for all } w' \textit{ such that } wRw'.$$

A first conceptual observation is that, under this semantics, OA means that playing a rational strategy entails A. In other words, obligations give *necessary* conditions for rationality. No profile is rational *unless* it satisfies A. The formula PA, on the other hand, here simply means that playing a rational strategy does not rule out A.

This interpretation of O has the welcome consequence of making the implication behind Ross's paradox, one of the classical problems of standard deontic logic, philosophically less problematic. The semantics above of course validates

$$OA \rightarrow O(A \vee B).$$

If we read OA in terms of necessary conditions for rational play, however, the antecedent of this formula says that no profile that makes A false is rational. Then, however, it follows that no profile that makes both A and B false is rational either.

This semantics yields a normal modal logic of rational recommendations (cf. [Blackburn *et al.*, 2002] for the general definition of normal modal logics). It validates the distribution of obligation over the material implication, i.e. the "K axiom":

$$O(A \rightarrow B) \rightarrow (OA \rightarrow OB)$$

and the so-called necessitation rule:

$$\text{From } A \text{ infer } OA$$

which says that all logically valid/provable formulas are obligatory. It should be noted that the game-theoretic interpretation has little to do with this. It is rather a consequence of the decision to construct the model $\mathcal{M}_\mathcal{G}$ as above. As we will see later on, the game-theoretic interpretation supports other modeling choices, some of which will not yield a normal modal logic for O.

Unlike standard deontic logic, however, this logic does not rule out normative conflicts, which in technical terms means that it invalidates the so-called "D axiom." Whether it does so depends in part on the underlying solution concept used to construct the models. In finite games, the set SD^ω of IESDS strategies is never empty, for instance. So for a given game $\mathcal{G}$, if we construct the model $\mathcal{M}_\mathcal{G}$ in such a way that $SD^\omega = \mathbb{S}(\mathcal{G}) = \{\sigma(w') : wRw'\}$ for all $w \in W$, then the set of states accessible from any state w will never be empty, which means that the D axiom will be valid in that model. On the other hand, it is well known that some games have no pure-strategy Nash equilibria. So D will not be valid if we take the latter to be our underlying solution concept. In those cases, however, normative conflicts will lead to what has been called deontic explosion: everything (and its contrary) becomes obligatory. This simple logic of rational recommendation in games does not rule out normative conflicts, but does not handle them well, either. For that, one would have to move to more sophisticated systems (cf. [Goble, 2014]). What the resulting logic of rational obligations in games would be is, as of now, an open question.

In cases where D is valid, this first attempt at capturing the deontic logic of rational recommendation is just standard deontic logic (SDL), as far as the non-iterated fragment is concerned.[1] The interest of this game-theoretic interpretation of SDL is that it provides concrete models to test our intuitions regarding known "paradox" or counter-intuitive consequences of that logic. As we saw, the Ross paradox appears less problematic. One can, however, easily generate a game-theoretic version of the contrary-to-duty paradox: Let us look back at the game we considered earlier. We saw that (U, L) is the unique IESDS solution of that game. One arrives at it in two steps, first eliminating R for Bob and then D for Ann. Observe, however, that if Bob were to play R, Ann's best response would be D. Now consider the following statements:

1. It ought to be the case that Ann plays U and Bob plays L: $\mathrm{O}(U \wedge L)$.

2. It ought to be the case that if Bob plays L, Ann plays U: wide scope: $\mathrm{O}(L \to U)$; narrow scope: $L \to \mathrm{O}U$.

3. It ought to be the case that if Bob plays R, Ann plays D: wide scope: $\mathrm{O}(R \to D)$; narrow scope: $R \to \mathrm{O}D$.

[1] The way we defined the relation R makes it a transitive and Euclidean relation, and so O is a K45 modality. These additional validities, however, only affect the iterations of deontic operators.

4. Bob plays R: R.

As usual in the contrary-to-duty case, in no combination are these four statements, in either narrow- or wide-scope form, both consistent and logically independent. Indeed, 1. entails both forms of 2. as well as the wide-scope reading of 3. Statements 1. and 4. are furthermore inconsistent with the narrow-scope reading of 3.

This first attempt at spelling out a deontic logic of rational recommendation is thus no better than standard deontic logic at handling conditional rational recommendations in games. Again, this is not a problem of the game-theoretic interpretation in itself, but rather of the modeling choices that were made in constructing this first semantics. Other choices are possible. See, for instance, [Tamminga, 2013], for a thorough treatment of such recommendations in games using contrary-to-duty conditional norms.

This simple logic also lacks the expressive power to distinguish between different rational recommendations stemming from different solution concepts. Some of that power can be recovered by adding propositional constants for strategies or strategy profiles and operators to describe the agent's preferences and/or beliefs, although it turns out that many such solution concepts can be subsumed under a single, relatively simple logic (cf. [Bjorndahl *et al.*, 2017] for details).

5 Alternative 2: Optimal and best

Apostel [Apostel, 1960] and more recently Tamminga in [Tamminga, 2013] have used a different approach to modeling rational recommendations in games. Although the implementation details differ, they share the basic idea: if there is more than one rational profile for a given game, all of them are rationally permissible but none is obligatory. Obligations only arise when the players have no alternative, i.e., where there is a *unique* solution to the game. Obligations, in that sense, are "uniquely action-guiding" [Tamminga, 2013], but permissions are not.

The following distinction is useful to spell out this idea formally:

Definition 5.1. *Let $\mathcal{G}$ be a game in strategic form and $\mathbb{S}(\mathcal{G})$ a solution concept.*

- *A profile σ is* optimal *whenever σ is in $\mathbb{S}(\mathcal{G})$.*
- *A profile σ is* best *whenever σ is the unique element of $\mathbb{S}(\mathcal{G})$.*

Apostel's and Tamminga's approaches answer Question 1 above in the same way as in the previous section. The difference lies in the answer to Question 2, namely in the interpretation of the deontic modality: they make all optimal profiles permitted, but only the maximal one, if any, is obligatory.

Definition 5.2. *Let $\mathcal{M}_\mathcal{G}$ be as in Definition 4.1. Then*

$$\mathcal{M}_\mathcal{G}, w \models \mathrm{P}A \quad \textit{iff} \quad \mathcal{M}_\mathcal{G}, w' \models A \ \textit{ for all } w' \textit{ such that } wRw';$$

$$\mathcal{M}_\mathcal{G}, w \models \mathrm{O}A \quad \textit{iff} \quad \textit{for all } w' \textit{ such that } wRw', \sigma(w') \textit{ is best and } \mathcal{M}_\mathcal{G}, w' \models A.$$

This alternative semantics for O and P differs substantially from standard deontic logic. First of all, P is a "box" here: PA requires *all* ideal states to satisfy A. In fact, P has here the same semantic clause as the O modality in our first semantics. So in this approach it is permission, instead of obligation, that provides necessary conditions for rationality. This entails, in particular, that permission distributes over conjunction, instead of the standard SDL principle of distribution over disjunction:

$$\mathrm{P}(A \wedge B) \ \leftrightarrow \ \mathrm{P}A \wedge \mathrm{P}B.$$

The operator O, on the other hand, is not a normal modality because it doesn't validate the necessitation rule:

$$\models A \quad \not\Rightarrow \quad \models \mathrm{O}A.$$

The reason for this is that when there is more than one solution to a game then no profile is best, falsifying the truth condition for OA. In those cases all formulas of the form OA turn out false, including those where A itself is a tautology. The operator O, however, inherits closure under conjunction from P, and so we do get that O is a regular modality:

$$\mathrm{O}(A \wedge B) \ \leftrightarrow \ \mathrm{O}A \wedge \mathrm{O}B.$$

Furthermore, P is no longer the dual of O here. $\neg\mathrm{P}\neg A$ only entails the existence of an optimal profile σ that makes A true. This of course neither ensures that *all* optimal profiles make A true nor that σ is the *unique* such profile. In other words, $\neg\mathrm{P}\neg A$ entails neither OA nor PA, as it does in SDL. The implication from $\neg\mathrm{O}\neg A$ to PA fails as well, this time owing to the fact, again, that P is a box modality.

This semantics, like SDL, does validate the standard form of the D axiom:

$$\mathrm{O}A \rightarrow \mathrm{P}A.$$

However, *unlike* with SDL, this does not entail that obligations are consistent, because PA is not equivalent to $\neg\mathrm{O}\neg A$. For the same reasons as for O in our original semantics, P allows for deontic explosion, and when this happens it will be inherited by the obligation operator as well.

This semantics thus gives rise to a non-normal obligation operator whose most notable feature is that as soon as the set of rational solutions for a game contains more than one profile there will be no rational recommendations (in the sense of true formulas of the form OA) to the players. In that case, the semantics will yield a number of rational permissions, but no rational recommendations, not even rational prohibitions, although some propositions will turn out *not* to be permitted. Just as permission here is not equivalent to the absence of prohibition, prohibition is also not equivalent to the absence of permission. So this semantics allows for cases where $\neg\mathrm{P}A$ is true but $\mathrm{O}\neg A$ is false.

Game-theoretically this semantics is very parsimonious when it comes to rational recommendations. Whatever solution concept is used, very few games have a unique rational solution. So in *most* cases this semantics will generate no obligations whatsoever to the players. Why adopt this semantics then? Tamminga [2013] correctly points out that in this semantics obligations are *uniquely* action-guiding: when a game has a unique solution then there is a determinate rational recommendation about what to do. However, determinacy and guidance seem different, and it is not clear why one should tie only the former to obligations. When the set of rational solutions is a strict subset of the set of all strategy profiles, or of the set of strategies for each agent, then some options are ruled out, that is, they are inadmissible from the perspective of that solution concept. In those cases, the players may not have determinate rational recommendations regarding a specific strategy to choose, but they do have rational recommendations about what to avoid, and that is a form of guidance.

6 Alternative 3: Permissions as sufficient conditions

I now turn to a third family of approaches for interpreting rational recommendations in games. Their most notable feature is that they interpret permissions in terms of sufficient conditions for rationality. These

approaches keep the same model construction and semantic clause for O as in Section 3. So obligations still give necessary conditions for rationality. The difference is in the clause for P:

$$\mathcal{M}, w \models \mathrm{P}A \quad \text{iff} \quad wRw' \text{ for all } w' \text{ such that } \mathcal{M}, w' \models A.$$

This inverted semantics—requiring all A-states to be ideal instead of all ideal states to be A-states—is well known by logicians under the name of "window modality" [Blackburn *et al.*, 2002]. As an interpretation for permissions in deontic logic, it has already been proposed by van Benthem in [van Benthem, 1979], but see also [Trypuz and Kulicki, 2013; Van De Putte, 2017] for recent studies and uses of this "deontic sufficiency" operator. As a model of rational permissions in *games*, this was already described in [Apostel, 1960], but has more recently been advocated in [Dong and Roy, 2015].

Like in the previous section, O and P are not dual in this new interpretation. Recall that $\neg\mathrm{O}\neg A$ is the dual of PA in SDL. It gets an existential or "diamond" interpretation. In other words, all that is required in SDL for PA to be true at a state w is that *some* A-states are accessible from w. This is perfectly compatible with cases where not all such A-states, or even very few of them, are accessible. In our game-theoretic interpretation this would mean that not all or even only very few of the rational profiles that are played at accessible states make A true. This is precisely what is excluded by the semantic clause of the window modality. PA now requires that *all* A-states are accessible from w. So even if $\neg\mathrm{O}\neg A$ is true, which means that some accessible states make A true (i.e., not all of them make A false), it is still possible that some non-accessible states also satisfy A. In such cases the semantics of the window modality will return PA as false. The converse direction of duality, i.e., from PA to $\neg\mathrm{O}\neg A$, does hold whenever the set of rational profiles is non-empty, but can fail otherwise. In those cases, the only permission that holds is P$\bot$, while obligations explode: we have OA for all A, so $\neg\mathrm{O}A$ for none of them.

Unlike in the previous section, however, we do not have that OA implies PA. This follows from the fact that the set of permitted propositions at each state w is the set of definable *sub*sets of the set of ideal states, while the set of mandatory ones is the set of definable *super*sets of the set of ideal states. This also makes sense intuitively: except in degenerate cases where only tautologies are obligatory, necessary conditions for rationality will not be sufficient ones as well.

As the previous remarks suggest, O is still a normal modality in this semantics, but P is not. The set of validities for these two modalities,

in a language containing the universal modality $\Box$ as well, is completely axiomatized by the axioms and rules in Table 1.

(K-$\Box$)	$\Box(A \to B) \to (\Box A \to \Box B)$	(Incl)	$\Box A \to \mathrm{O}A$
(K-O)	$\mathrm{O}(A \to B) \to (\mathrm{O}A \to \mathrm{O}B)$	(CoIncl)	$\Box\neg A \to \mathrm{P}A$
(OR)	$\mathrm{P}A \wedge \mathrm{P}B \to \mathrm{P}(A \vee B)$	(WP)	$(\mathrm{O}A \wedge \mathrm{P}B) \to \Box(B \to A)$
(NEC)	$\dfrac{\vdash A}{\vdash \Box A}$	(Flip)	$\dfrac{\vdash A \to B}{\vdash \mathrm{P}B \to \mathrm{P}A}$

Table 1: The complete axiom system for permission and obligation as necessary and sufficient conditions, respectively, for rationality. The $\Box$ operator is an S5 normal modality.

Among these axioms, (WP) is notable for expressing the logical relationship between necessary and sufficient conditions for rationality. Indeed, if B is rationally permitted, then it is sufficient for a rational play. That A is obligatory, on the other hand, means that no profile is rational unless it is played in an A-state. But then all B-states must be A-states, which is precisely what the consequent of (WP) says.

The most notable feature of this axiomatization, however, is the Flip rule, which in the present context is provably equivalent to the Free Choice Permission principle:

$$\mathrm{P}(A \vee B) \to \mathrm{P}A \wedge \mathrm{P}B.$$

This principle is infamous, among other things, for trivializing standard deontic logic and causing further well-known paradoxes [Hansson, 2013]. Most of the damage done in SDL is prevented here by the fact that P is neither the dual of, nor implied by O. Some apparently counterintuitive consequences remain, however. Substitution under logical equivalence indeed gives the following provably equivalent formulation of Flip or Free Choice:

$$\mathrm{P}A \to \mathrm{P}(A \wedge B)$$

for any formula B whatsoever. This gives rise to the well-known "Vegetarian Free Lunch" example [Hansson, 2013], where a permission to order a vegetarian lunch entails a permission to order that same lunch but not pay for it.

This type of example is less poignant here than in SDL. If any formula A is permitted here, this means that *all* cases where that formula

is true are cases of rational play. Of course one can form the conjunction of such a formula with another, say B, for which we have $\neg \mathrm{P}B$, and get $\mathrm{P}(A \wedge B)$. But observe that $\neg \mathrm{P}B$ is consistent both with $\mathrm{O}\neg B$ and with $\neg \mathrm{O}\neg B$. In the latter case, this means that while B is not strong enough to ensure a rational play, it is nonetheless consistent with it. In those cases $\mathrm{P}(A \wedge B)$ is unproblematic. If, however, we have $\mathrm{O}\neg B$, then (WP) yields directly $\Box(B \to \neg A)$, which just says that A and B are mutually exclusive. But then the permission $\mathrm{P}(A \wedge B)$ is equivalent to $\mathrm{P}(\bot)$, i.e., to a permission to do the impossible. Although admittedly not helpful as a guide to decision-making, this permission is otherwise philosophically harmless.[2]

I finish this section with a short mention of a close cousin of this logic, developed in [Anglberger *et al.*, 2015] and further studied in [Van De Putte, 2016]. The semantics for this logic turns out to be rather different than the ones we covered so far, because neither O nor P are normal modalities in this system. So instead of a Kripke semantics in the style of the previous sections, [Anglberger *et al.*, 2015] use a natural generalization called neighborhood semantics [Pacuit, 2017]. The basic idea behind that system, however, is simple: like in the alternative just presented, permissions are sufficient conditions for rationality, but obligation is defined as the *unique* "weakest permission", which in our deontic interpretation just boils down to the set of states that get assigned rational profiles.

The resulting logic is surprisingly close to the one presented in Table 1. It includes the same axioms for P, and the (WP) interaction axiom, but augments these with the principle that obligation implies permission:

$$\mathrm{O}A \to \mathrm{P}A.$$

Of course, in the presence of this axiom, O cannot be a normal modality anymore, on pain of recovering the usual trivialization generated by adding the Flip rule to SDL. Instead, we get that obligations become unique up to extensional equivalence, as expressed by the following consequence of (WP) together with the implication from O to P:

$$\mathrm{O}A \wedge \mathrm{O}B \to \Box(A \leftrightarrow B).$$

In the present context, this boils down to saying that the players have only one obligation, again up to extensional equivalence, namely to play

[2] For an alternative solution to this problem, this time using truthmaker semantics, see [Anglberger and Korbmacher, 2020].

a rational profile. Obligation is then axiomatized by replacing necessitation and the K axiom with the same extensionality rule, together with the two interaction principles (WP) and "O implies P".

7 Conclusion

The goal of this chapter was twofold: first, to introduce the reader to applications of deontic logic to game theory, and second, to emphasize that there is no unique way of representing obligations and permissions stemming from one particular solution concept: many modeling choices have to be made. I have presented three of them, highlighting the fact that each comes with specific philosophical pluses and minuses.

The chapter has been primarily targeted at deontic logicians interested in game theory. There are many reasons why a game-theoretic interpretation may be of interest to them. A prominent one is that it provides a concrete interpretation against which one can test general intuitions about, for instance, what counts as a paradoxical or counter-intuitive deontic principle. We saw an example of that with Ross's paradox, which appears less problematic once O and P are respectively interpreted as necessary and sufficient conditions for rationality. Another reason for a deontic logician to be interested in game-theoretic models is that some natural interpretations of O and P, for instance, in terms of "best" and "optimal" strategies, give rise to logical systems that have up to now not been thoroughly studied.

It is, however, also legitimate to ask why game theorists might be interested in deontic logic. First and foremost, the considerations above reveal that the pre-theoretical idea of what game-theoretic rationality prescribes is in need of an explication, and that even basic tools from deontic logic can be of help in providing one. Second, the standard logical results, especially concerning axiomatization and definability, shed light on the core commitments behind such conceptual explications. Complexity results for satisfiability or model-checking could also be compared to existing ones regarding the difficulty of computing solutions like IESDS or Nash equilibria.

It should be noted that the general project of circumscribing the logic of rational prescriptions is not limited to games or the theory of strategic interaction. Some of the solution concepts mentioned here, for instance IESDS, are grounded in a decision-theoretic understanding of rational decision making under uncertainty [Pacuit and Roy, 2017]. As such there should be substantive overlaps between the logics of ratio-

nal recommendations in games and those of decision under uncertainty more generally. This is already visible in [Bjorndahl *et al.*, 2017]. More work should be done, however, to unify deontic logics for games and for individual decisions, e.g., those in [Cariani, 2016] or [Horty, 2001].

Of course, many questions are left open, even if we restrict our inquiry to the specific applications to game theory that I have surveyed. The most urgent one concerns the study of rational recommendations stemming from specific solution concepts. As mentioned, this would require following steps taken long ago in deontic logic and studying the combination of obligation and permission with temporal, agentive, preferential or epistemic notions. Such languages of course exist, and many "modal characterizations" [van der Hoek and Pauly, 2007; van Benthem and Klein, 2020] of solution concepts can be readily given a deontic rider. The task, in those cases, is thus just as much to explore the mathematics of these systems as to assess their conceptual and philosophical import. Another emerging topic is the interaction between individual and group rationality in games. Recent advances on that have been made in [Tamminga and Duijf, 2017; Tamminga and Hindriks, 2019], but more needs to be done for specific solution concepts, and for looking beyond one-shot, simultaneous games. Finally, this chapter has not touched at all on questions of definability and the relative expressive powers of the different languages and semantics introduced here.

References

[Anglberger and Korbmacher, 2020] A. Anglberger and J. Korbmacher. Truthmakers and normative conflicts. *Studia Logica*, 108(1):49–83, 2020.

[Anglberger *et al.*, 2015] A. Anglberger, N. Gratzl, and O. Roy. Obligation, free choice, and the logic of weakest permissions. *Review of Symbolic Logic*, 8(4):807–827, 2015.

[Apostel, 1960] L. Apostel. Game theory and the interpretation of deontic logic. *Logique et Analyse*, 3(10):70–90, 1960.

[Bjorndahl *et al.*, 2017] A. Bjorndahl, J. Halpern, and R. Pass. Reasoning about rationality. *Games and Economic Behavior*, 104:146–164, 2017.

[Blackburn *et al.*, 2002] P. Blackburn, M. De Rijke, and Y. Venema. *Modal logic*, volume 53. Cambridge University Press, 2002.

[Cariani, 2016] F. Cariani. Deontic modals and probabilities: One theory to rule them all? In N. Charlow and M. Chrisman, editors, *Deontic modality*, pages 11–46. Oxford University Press Oxford, 2016.

[De Bruin, 2010] B. De Bruin. *Explaining games: The epistemic programme in game theory*, volume 346. Springer Science & Business Media, 2010.

[Dong and Roy, 2015] H. Dong and O. Roy. Three deontic logics for rational agency in games. *Studies in Logic*, 8(4):7–31, 2015.

[Goble, 2014] L. Goble. Prima facie norms, normative conflicts, and dilemmas. In D. Gabbay, J. Horty, X. Parent, R. van der Meyden, and L. van der Torre, editors, *Handbook of Deontic Logic and Normative Systems*, pages 241–352. College Publications, 2014.

[Hansson, 2013] S.O. Hansson. The varieties of permissions. In D. Gabbay, J. Horty, X. Parent, R. van der Meyden, and L. van der Torre, editors, *Handbook of Deontic Logic and Normative Systems*, volume 1. College Publication, 2013.

[Horty, 2001] J. Horty. *Agency and Deontic Logic*. Oxford University Press, 2001.

[Osborne and Rubinstein, 1994] M.J. Osborne and A. Rubinstein. *A course in game theory*. MIT press, 1994.

[Pacuit and Roy, 2017] E. Pacuit and O. Roy. Epistemic foundations of game theory. In E.N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, summer 2017 edition, 2017.

[Pacuit, 2017] E. Pacuit. *Neighborhood semantics for modal logic*. Springer, 2017.

[Tamminga and Duijf, 2017] A. Tamminga and H. Duijf. Collective obligations, group plans and individual actions. *Economics & Philosophy*, 33(2):187–214, 2017.

[Tamminga and Hindriks, 2019] A. Tamminga and F. Hindriks. The irreducibility of collective obligations. *Philosophical Studies*, pages 1–25, 2019.

[Tamminga, 2013] A. Tamminga. Deontic logic for strategic games. *Erkenntnis*, 78:183–200, 2013.

[Trypuz and Kulicki, 2013] R. Trypuz and P. Kulicki. On deontic action logics based on boolean algebra. *Journal of Logic and Computation*, 25(5):1241–1260, 2013.

[van Benthem and Klein, 2020] J. van Benthem and D. Klein. Logics for analyzing games. In E.N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, summer 2020 edition, 2020.

[van Benthem, 1979] J. van Benthem. Minimal deontic logics. *Bulletin of the Section of Logic*, 8(1):36–42, 1979.

[Van De Putte, 2016] F. Van De Putte. Obligation as weakest permission: A strongly complete axiomatization. *Review of Symbolic Logic*, 9(2):370–379, 2016.

[Van De Putte, 2017] F. Van De Putte. "that will do": Logics of deontic necessity and sufficiency. *Erkenntnis*, 82(3):473–511, 2017.

[van der Hoek and Pauly, 2007] W. van der Hoek and M. Pauly. Modal logic for games and information. In P. Blackburn, J. van Benthem, and F. Wolter, editors, *Handbook of Modal Logic*, volume 3, pages 1077–1148. Elsevier, 2007.

Olivier Roy
Universität Bayreuth, Institut für Philosophie, Germany
Email: olivier.roy@uni-bayreuth.de