1
Ethical Learning, Natural and Artificial

Peter Railton

1.1. Introduction

There is no shortage of urgent ethical questions about the responsible development and deployment of artificial intelligence. Artificial intelligence is a *fundamental* technological innovation in the sense that, besides adding new technological possibilities of its own, it alters the capabilities and potential benefits and risks of a wide range of other technologies, including such "soft technologies" as social practices and institutions. One has only to imagine an area of human life—work, communication, governance, mobility, medicine, warfare—in order to have exciting *and* disturbing potential effects of artificial intelligence spring to mind, which grow more exciting and more disturbing the further one imagines artificial intelligence to be capable of developing.

This paper will not address directly any of these particular problems about the responsible development and deployment of artificial intelligence, important as they are; rather it considers a question that could be relevant to all of them, since it concerns the nature of artificial intelligence itself. Increasingly, artificial systems will be exercising life-affecting functions or making life-affecting decisions—in piloting a vehicle, in home healthcare, in hiring and firing, in monitoring and shaping the information we receive—that we would not ordinarily entrust to someone lacking in sensitivity to ethical concerns. How, then, might artificial systems come to be appropriately sensitive to ethical concerns? Moreover, how might such sensitivity be a core part of their intelligence and capacities? My primary focus will be on how this might be possible. To a first approximation, we can characterize sensitivity to ethical concerns as a robust, reliable capacity to detect and respond appropriately to ethically relevant features of situations, actions, agents, and outcomes. Our answer to questions about how we can responsibly develop or deploy artificial systems will depend significantly upon the extent to which such apt responsiveness to ethically relevant features is possible.

A closely related question, to my mind, is this: As artificial intelligence becomes more general and capable, it will give rise not only to new technological possibilities, but to new classes of *agents* that operate independently of direct

Peter Railton, *Ethical Learning, Natural and Artificial* In: *Ethics of Artificial Intelligence*. Edited by: S. Matthew Liao, Oxford University Press (2020). DOI: 10.1093/oso/9780190905033.003.0002

human supervision or control, and with which we can increasingly have *social* rather than merely instrumental relations. For example, as a matter of safety, it may be important for artificial agents to be able to refuse to comply with certain commands, or to have an element of uncertainty about the goals we give them, so that they pay attention to accumulating evidence of harms, bias, or dysfunction and can make their own decision to suspend pursuit of such goals or seek more information and advice. While there are dangers inherent in creating highly capable artificial agents with enough autonomy to question the goals they are given on grounds of harm, bias, or dysfunction, there is greater danger in creating highly capable artificial agents lacking any capacity to do so. Think only of the same issue raised with respect to raising human (and presumably highly capable) agents; as we will see, human infant ethical development typically proceeds in a sufficiently autonomous way that three- to four-year-olds will question rules given to them by persons in authority when they believe these rules to cause inappropriate harm or be unfair.[1] Moreover, often the information artificial agents need to make such decisions will be available most rapidly or reliably from other artificial agents. We will need to find ways to coordinate, cooperate, collaborate, and compete peacefully and productively with artificial systems, seen as independent parties whose behavior we cannot simply dictate. Moreover, artificial systems will need to find ways to coordinate, cooperate, collaborate, and compete peacefully and productively with us in return, and with each other. Artificial systems capable of projecting and evaluating future courses of action, of assessing benefits and harms to self and other, of making commitments, and of regulating their own behavior accordingly will be capable of something like social-contract reasoning: we could negotiate with them terms of mutually beneficial cooperation that all of us would constrain ourselves to follow.

This capacity for social-contract reasoning, and for mutual constraint for mutual benefit, does not presuppose a capacity for qualitative experiences or emotions akin to humans. In humans, our actively norm-governed life together is greatly enhanced by our capacity for a range of affective states—empathic simulation and emotions such as loyalty, guilt, forgiveness, and so on[2]—but an actively norm-governed life does not seem to require such feelings, so long as there are sufficiently developed agential capacities for self-regulation, representation of others' goals and information, and the formation of conventions, agreements, or commitments. Even before we have to contend with possible "super-intelligences,"[3] we will need to ask how to contend with artificial agents from whose capacities we could greatly benefit but whose cooperation with us will not be entirely up to us and may depend upon negotiation in which we seek to find common ground for working together and according to each other's goals. Indeed, for intelligent systems to be able to robustly and reliably detect and respond to ethically relevant features they may need to have at least this much

autonomy in deciding whether to work with the particular human or artificial agents who might seek to control them, and for what purposes.

We are not without experience of highly capable nonhuman agents lacking a unified consciousness or affective states but possessing extraordinary levels of information and problem-solving ability, and whose aims may differ from our own in ways that require us to negotiate with them if we are to gain the benefits they make possible. Corporate entities—governments, corporations, universities, institutes, unions, political parties—can have a distinctive set of goals related to their own purposes or conditions for survival and flourishing, which may overlap with but also fail to be the same as those of the individuals who compose them or are affected by them. They possess capacities for pursuing values and holding themselves to norms, for future projection and planning, for entering into (or failing to enter into) cooperative arrangements, strategic alliances, mutual commitments or contracts, and for incurring, carrying out, and policing compliance with associated obligations. At the same time, they are not fully transparent in their inner processes; asking how an action by a corporate entity came to be taken may not yield a determinate decision-process with clear lines of responsibility. Asking how we might enter into mutually beneficial, mutually constrained, normatively governed relations with emerging agents possessing higher-than-human intelligence is like asking how we are able to enter into such relations with governments, corporations, and so on. We have made a fair amount of progress in developing countervailing institutions and normative practices that enable us to work with such agents in ways that can be mutually beneficial. But this is still a work in progress.

1.2. A Social Perspective

Artificial neural networks, I am told, were originally inspired by the thought that naturally occurring cortical architecture is the result of countless generations of selection for a capacity to learn and act intelligently, and so is a plausible basis upon which to build artificial intelligence. Now that artificial neural networks have become sufficiently deep and fast, and data have become sufficiently plentiful, this inspiration is bearing fruit. Perhaps, as we look forward to the development of *general* artificial intelligence, we should look for inspiration at the most distinctive characteristics of the naturally occurring creatures that appear to have achieved the highest levels of general intelligence: humans. And humankind is at least as distinctive, relative to the wider animal world, for our *social* capacities as for our sheer intellect. Many animals, of course, live in complex social groups, but humans are unusual in the extent of their capacity for large-scale coordination and cooperation with nonkin, open-ended exchange of information,

and normative self-regulation in light of long-term, abstract impersonal goals. If one thinks of general intelligence as a capacity for open-ended problem-solving, then our capacities for building and sustaining shared practices to solve shared problems are central to our general intelligence. These capacities enable us to create *epistemic* as well as ethical communities, leveraging our individual abilities in ways that can carry us far beyond anything we could accomplish as individual agents or inquirers.

This paper is an exploration of the idea that the project of building highly effective, generally intelligent *artificial* epistemic agents should be seen as connected with building artificial agents capable of apt responsiveness to ethically relevant features. This idea has a certain advantage in thinking about ethics and artificial intelligence. If we imagine that achieving responsiveness to ethically relevant features in an artificial system is a matter of *adding* a novel capacity or set of principles to an already fully formed general intelligence, then it might also be imagined that this capacity or these principles could readily be *subtracted* from such an intelligence without cognitive loss. If, instead, there is a root connection between full development of the capacity to be appropriately responsive to epistemically relevant features and full development of the capacity to be appropriately responsive to ethically relevant features, then responsiveness to ethically relevant features could be a *deep* feature of artificial systems with high general intelligence and problem-solving ability, not easily removed without serious impairment of other aspects of general intelligence and problem-solving. Suggestive evidence comes from the way in which some psychological disorders that seem to impair the development of appropriate responsiveness to ethically relevant features, such as psychopathy, tend also to have costs to the full development of more general human intelligence, understood as problem-solving ability.[4] Contrary to the popular idea of the psychopath as the height of rationality, intelligence, control, and savvy is the research indicating that psychopaths show serious deficits in attention, impulse-control, and ability to accurately represent likely negative future outcomes—for themselves as well as others. These deficits then help explain the difficulty of psychopaths in holding themselves to long-term goals, plans, or relationships.[5] But the point is not solely about individuals. Persons with "Machiavellian" personal disorders may be quite intelligent by conventional measures and can achieve considerable long-term success by taking advantage of the vulnerabilities of ordinary agents and practices,[6] but it is one thing to be able to exploit an existing epistemic community and another thing to be able to build and sustain one that is as effective as possible at gaining knowledge. The latter is closer to the ultimate goal of artificial intelligence research. For example, individuals scoring high in Machiavellian psychological profiles are more likely to adopt an "economically rational" strategy in trust games, whereas no known human community has this as the predominant

disposition,[7] and the communities, small and large, that most effectively work together to enlarge their capacities depart the most from this strategy.[8] Imagine the difference in learning to drive safely in an open-ended array of situations if a community of self-driving cars shares individual driving data rather than each using the data it acquires to gain whatever strategic advantages it can over the others.

It is important to distinguish our question about responsiveness to ethically relevant features from asking how artificial systems might come to have the distinctive qualitative experiences or affective responses of typical human moral agents. For example, empathic simulation appears to play an important role in helping humans to understand one another,[9] but an artificial system can engage in empathic simulation without "reliving the experience" of others if artificial systems can become sufficiently skilled at modeling others' internal states on the basis of observed behavior and can accord intrinsic weight to others' imputed goals or utility functions in evaluating simulated courses of action during decision-making.

Consider by analogy the fact that developing artificial systems capable of being effectively responsive to a range of semantically relevant features in natural language need not await the development in such systems of the full range of human thought and feeling. We might think that no artificial system could grasp the full meaning of "Where are the snows of yesteryear?" without feeling a pang of nostalgia, but a genuinely intelligent artificial system might nonetheless be able to represent the essential semantic features of this sentence and to capture enough of the pragmatics of English usage to give it as a suitable English translation of the original, "Où sont les neiges d'antan?," rather than, say, the flat-footed "Where are the snows of previous years?" Moreover, and importantly for our purposes, recent developments in artificial natural-language processing suggest that artificial systems may be able to *learn* underlying syntactic and semantic structures of language from the task of developing compact, hierarchical, predictive, or generative models of large bodies of linguistic data and can use these models to guide interpretation.[10] Unlike systems that are preprogrammed with grammatical information, these systems use fairly generic learning methods to "acquire" from exposure to language data latent structures *in* language that then can be used for tasks like interpretation, translation, and similarity judgments for an open-ended array of sentences. Might similar kinds of general-purpose learning capacities enable artificial systems to extract from the context of human interaction ethically relevant latent structures of situations, actions, agents, and outcomes? And might this in fact be much closer to the way actual humans become sensitive to ethically relevant features or make ordinary, "intuitive" ethical judgments? This brings us to a developmental perspective on human cognitive and ethical capacities.

1.3. A Developmental Perspective

Recent years in developmental psychology have seen the emergence of learning-based, "constructivist" or "theory forming" approaches to cognitive phenomena previously attributed to specialized "innate modules."[11] For example, in the case of language learning, it has long been recognized that infants receive relatively little explicit instruction in language in their early years, yet during these years normally developing children acquire a remarkable degree of fluency in understanding and producing an open-ended array of novel sentences in their native tongue. Positing an innate, generative language module seemed to offer the only explanation of how this could occur, given the limitations of "associative learning" and the disproportion between the finite amount of language and language training to which children are typically exposed and the open-ended competence they acquire. This is sometimes known as the "poverty of the stimulus" argument.[12] But *is* the stimulus really impoverished? And is "associative learning" really so limited?

Since the heyday of innatism, we have learned a considerable amount from cognitive science about the potential for experience-based learning of rich, hierarchical structures, and from developmental psychology about the highly active experiential life of infants, even in their earliest days and weeks.[13] Moreover, innatism always faced the problem that infants must somehow already be able to detect many structural and contentful features of language in order to *apply* a category- and rule-based "innate grammar." After all, infants begin life in a complex, continuous acoustic environment within which they must learn to distinguish overheard *language* from other elements in the continuous stream of sounds and noise, and to attend to overheard language closely enough to detect patterns that permit the extraction of discrete, recurring units and combinations in the language, despite, for example, the wide variation in the acoustic profile of individual voices. Moreover infants need to be able to detect signs of adult attention and to track the intended referents of adult gestures or words. These are already formidable learning tasks in modeling structural features of the world the infant inhabits, and they must be solved for her specific acoustic and social environment. An innate grammar module on its own would not equip the infant to accomplish them. What could?

Recent developments in machine learning applied to natural language have begun to suggest how such learning might be possible through probabilistic means, even in the absence of much by way of explicit linguistic instruction.[14] Infants are, after all, exposed to a very large amount of overheard language and have at their disposal a very large amount of fast, flexible computational capacity. It seems they put both of these to good use. For example, we now have evidence that infants in the first weeks of life have begun to discriminate overheard speech

from the rest of their acoustic environment and are beginning to form calibrated expectations about phonetic regularities, which are manifest in greater surprise at, and interest in, novel or anomalous sequences of phonemes—a characteristic feature of probabilistic learning.[15] Over the course of the first year and a half, while an infant's explicit language capacity and adult explicit linguistic instruction are both typically limited, young infants have begun to piece together the social and intentional structure around them. By nine months they can discern others' goals on the basis of their behavior,[16] and by twelve to sixteen months they can relate means to goals,[17] follow others' attentional cues,[18] engage in joint attention,[19] and identify the intended referents of their words or gestures.[20] We see, then, a pattern of emerging competencies of a kind important for the development of language, accomplished gradually through the course of experience in a way that resembles their gradual learning of other kinds of causal relations and regularities in their world.

The stimulus infants receive, then, is not so impoverished after all, stretching over many months of observation of the behavior of persons and objects in their near vicinity. And probabilistic forms of learning turn this seemingly "passive" experience into more than "mere association." Instead it is a form of active *experimentation*, with the continuous formation of expectations on the basis of observed associations and continuous feedback from discrepancies between such expectations and actual outcomes. Since the physical and social world contain very significant structure, more effective and efficient prediction pushes infant learning in the direction of representing such structure, favoring the development of internal models that use abstraction and hierarchy to generalize projectively, without the need to posit an innate "language module."[21]

We can connect this idea of learning via experiential modeling to the child's challenge in moving from observation to action by reflecting on the so-called "Good Regulator Theorem" of control theory,[22] which holds that ideally effective and efficient regulation of a system requires the building and use in decision-making of a model of that system—a model representing the underlying structures and potentials of the system. Such a model can be used in a forward direction for intelligent simulation and action selection, and in an inverse direction for learning from subsequent experience. Models of this kind can also play a fundamental role in the development of motor control skills.[23]

Suppose, then, that we think of the infant mind as *regulating* its interactions with the environment, exercising whatever capacities it can to get its needs met. And no part of the infant's causal environment is more important for her than the *agents* in her life, so that causal and social learning are intimately linked, and intuitive psychology emerges alongside intuitive physics. It would, after all, be very difficult for the infant to build a predictive model of the world around her without taking into account the distinctive ways in which agents behave, and

beginning to model the "internal" as well as external sources of such behavior, much as infants begin to model latent as well as manifest causal relations.[24] Evidence suggests that infants develop piecewise an increasingly complex "theory of mind" or model of agents as continuing entities whose behavior is the product of perception, motivation, emotion, belief, and intention.[25]

Moreover, while the infant might start by using her own mind as a matrix for understanding others and their actions,[26] the pressure to develop more reliable expectations of others pushes in the direction of representing others' mental states in their own right—not as projections of the infant's own states. We see emerging in infants an ability to grasp that others may differ, first, in motivation, then in belief, then in perceptual knowledge, then in possessing false beliefs, and then in hidden emotions.[27]

Spatial representation in foraging animals (ourselves included) appears to involve the construction through experience of non-egocentric as well as egocentric maps.[28] These spatial representations can then be used to associate expected rewards with nonproximate locations and to simulate and compare possible pathways toward these rewards, facilitating more efficient and effective foraging.[29]

Likewise infant mapping of social space and its possibilities involves an ability to represent how things are in non-egocentric as well as egocentric terms, making possible more accurate, less position-dependent simulations of potential social interactions and evaluation of their likely outcomes. Over the course of the first years of life, when infants have only limited causal powers of their own, observation of others' actions and outcomes plays a fundamental role in the development of their own expectations and understanding.[30] The non-egocentric *epistemic evaluation* of others—observing others' interactions to map the reliability and competence of agents in their interactions with third parties—comes to play a critical role in shaping who infants are disposed to imitate or learn from, independent of personal affiliation.[31] As we will see, this ability to form and be guided by non-egocentric as well as egocentric representations and evaluations, which might be driven in the first instance by the need for accurate prediction, is of special interest for *ethical* development in children, since among the fundamental features of ethical evaluation are that it calls for an ability to represent non-egocentrically the nature and magnitude of the concerns of others, the likely results of one's own actions, the causal-intentional structure of others' actions, and whether others are reliable or trustworthy.

1.4. Default Trust and Default Cooperation

What are some of the characteristics of a developing psyche that would promote this kind of integrated learning about the causal world of things and agents?

Clearly, infants need to be motivated to attend carefully to experience and to notice patterns. They need to form expectations based upon such patterns, and to find failed expectations discomfiting in themselves, even when this does not directly touch their interests. And they need to respond to such anomalies by increasing their attention and effort, not by simply shrinking the scope of their expectations. This collection of features we can think of as *curiosity*, a form of internal motivation to learn above and beyond any more specific purpose the infant might have.

But curiosity is not enough. The brief description just given presupposes that infants are also disposed to *rely upon* or *trust* their own faculties—perception, association, memory, and so on—even without any guarantee of the reliability of these faculties. Without such a disposition toward *default reliance* or *default trust*, even an infant natively equipped with good eyes and ears and a keen mind would remain trapped in ignorance. After all, any evidence she might gather of the reliability or unreliability of her faculties would already depend upon the use of those faculties, in effect giving them some measure of default epistemic authority. Once some measure of default reliance or trust is in place, then the formation of expectations can begin to generate feedback from subsequent experience--a kind of bootstrapping. Bootstrapping does not mean *indefeasibility* however; indeed, default reliance and trust operate in the service of generating more determinate guesses, creating the potential for more informative errors and growing more selective or calibrated over time.

Somewhat metaphorically, we can think of such default, defeasible trust as a "prior" that enables the infant to *cooperate* with her faculties, by "playing" a cooperative move on the first turn by forming expectations as she would if her faculties were reliable, yet with no security that her faculties will prove cooperative in return by yielding reliable information. In contrast, for her to refuse to extend any unsecured cooperation to her faculties (in this metaphorical sense) would be a self-defeating epistemic strategy—not by incurring a risk of believing something false but by undermining the possibility of believing anything at all.

Consider now that portion of the infant's epistemic engagement with the world that is social, and where cooperation can be less metaphorical. Here too an infant initially disposed not to rely upon or trust those around her until she has confirmation that such reliance and trust will be well-placed would cut herself off from the very experiences she would need in order to learn whom or what to trust, and how much. As before, initial trust can be modulated by subsequent experience, so that expectations can become better calibrated to actual outcomes, and reliance and trust more selective.[32]

Infant default reliance and trust extend beyond the epistemic and can play a vital role in initiating cooperative relations with others. Early on, infant responsiveness *reinforces* adult attention, facilitating development of reliable channels

of communication between infants and caregivers that do not depend upon language. Infants are typically disposed to reciprocate care, to the extent that they can. For example, by the second year infants still crawling or toddling are able to form representations of adult goals from failed as well as successful adult behavior[33] and are spontaneously disposed to initiate an attempt to help an adult complete a failed task, even a stranger, and without encouragement or promised reward.[34] And toddlers who have participated in a successful shared task with a novel partner are spontaneously motivated to share the gains achieved by the task, again without explicit encouragement or reward.[35] These are manifest forms of a general disposition to default cooperativeness that has in fact been operative in the infant since early weeks of life, helping her to establish positive, reciprocal relations with those around her.

Infancy is an extreme case in which an individual's problem-solving capacities depend upon developing sustained, selective engagement with and reliance upon others. But as humans go through life, what they learn and what they are capable of achieving do not cease to depend extensively on coordination or cooperation with others. If anything, the scope of the coordination and cooperation with others needed for continued learning and success in attaining one's ends *grows* with time. For this to be sustainable, individuals must be motivated both to trust help from unrelated others and to help unrelated others in ways that reward their trust. As Hobbes pointed out over three hundred years ago, mutually beneficial cooperation among strangers is possible when individuals are disposed to initiate cooperation without requiring initial security (e.g., as a credible way of signaling willingness to cooperate) and to reciprocate cooperation when it is received.[36] More recently, game theorists have shown that this set of dispositions can become widespread within a population and be effective in resisting "invasion" by more opportunistic agents.[37] And a large-scale survey of hunter-gatherer societies suggests that a capacity for coordination and cooperation with others, including nonkin, mediated by forms of reciprocity that are indirect and temporally extended, may play a central role in explaining how human hunter-gatherers have succeeded over millennia in maintaining egalitarian social cohesion in the face of limited resources, without the forms of dominance hierarchy found in the great apes.[38] Recent research suggests that the disposition to give weight to the interests of others that is not simply mediated by one's own interests is something like the default stance of ordinary human interaction and can be self-reinforcing[39]—including, one might stress, human communication and information exchange, as the norms of conversation attest.

So much of what we think of as an individual human's general human intelligence or problem-solving capacity is really social in origin, character, or operation that we should think of the ability to initiate and sustain productive social connectedness with others as an additional basic faculty of learning, which

supplements other basic faculties like perception, memory, and reasoning. Default cooperative dispositions with respect to others therefore are as much a part of the human capacity for learning as default cooperative dispositions with respect to one's own basic mental faculties.

Language is fundamental to general intelligence and problem-solving capacities typical of humans, which are able to draw upon social knowledge built up over generations of experience. And language is an outstanding example of what default cooperative dispositions among nonkin can accomplish for any species that can achieve them. A shared language can be sustained only because enough speakers regularly use the language with sincere and helpful communicative intent to make it worthwhile for us to speak with each other and rely upon what each other says—to make openness to conversational exchange, overall, a positive-sum activity. Open conversational exchange among strangers is a form of mutual constraint and contribution for mutual benefit, and it plays an essential role in knitting together and facilitating the large-scale forms of cooperation and accommodation upon which human culture depends.

Individual human intelligence and problem-solving ability at age two is said to be quite comparable to that of a chimpanzee of the same age. But human two-year-olds are able to do something even adult chimpanzees are not, and that is fundamental to the extensive growth of human intelligence and problem-solving ability: to come together spontaneously with others to accomplish a task requiring joint attention and coordinated playing of understood roles, and, equally spontaneously, to share the rewards of cooperation with others without further incentive.[40] The divergence in cognitive accomplishment and practical problem-solving that comes as humans work together—the emergence of shared languages, of extensive forms of social learning, culture, and exchange—explains why *Homo sapiens* could overrun the planet, making their own habitats as needed, while *Pan troglodytes* is at risk of disappearing from the wild as its natural habitat shrinks.

1.5. Ethical Development and Ethical Judgment

We have spent so much time on the questions about the capacities underlying aspects of language and epistemic development because they afford us insight into the capacities that underlie ethical development as well. It is no accident that the norms of conversation, for example—of mutual recognition, of according others some authority to contribute, of seeking to determine the communicative intent of others and signaling this to them, of seeking to reply in ways that could be comprehensible, relevant, and responsive to others' concerns, and so on—are so close to norms for productive epistemic exchange. Now we will add: it

is no accident that they are also so close to norms for ethical interaction. Indeed a large and influential tradition in ethics, *communicative ethics*, is built around this fact.[41]

Intriguingly, the step-wise development of children's ability to model others' minds predicts a range of features of infant behavior that have strong relevance to ethical learning. For example, even controlling for other abilities, a child's development of theory of mind is predictive of her current and future *maturity*, as manifest in the ability to form positive relations with peers. Such abilities include: understanding the needs and interests of others, even when different from oneself or one's group; standing up for one's own opinions, needs, and rights; successfully joining new groups or welcoming new members into one's own group; playing or working together with peers without conflict; and coping with conflicts that do arise.[42] These are all skills that involve apt responsiveness to ethically relevant features, as understood by virtually any widely held ethical theory. Just as there was a parallelism in the development of causal understanding and theory of mind, there is a parallelism in the development of theory of mind and capacity to be aptly responsive to ethically relevant features.

For example, assessment of *intent* is a core component of understanding the causal, epistemic, and ethical character of an action, so acquiring the ability to distinguish intentional from unintentional actions is important for prediction (e.g., what to expect next), learning (e.g., whether an adult error was the result of ignorance or is a sign of unreliability), and ethical assessment (e.g., whether a harmful action by an individual was an accident or is a sign of ill will or untrustworthiness). By the end of their first year infants have begun to use situational cues to determine whether an action is intentional to modulate their responses in all three domains.[43]

Ethical development appears to begin earlier than the explicit inculcation of social norms by adults and also to develop in ways that are both more basic—for example, in grasping what behavior, in a given context, constitutes a harm—and more autonomous than external instruction. An example of autonomy, mentioned earlier, is the fact that three- and four-year-olds across a range of cultures show a spontaneous ability to question rules given to them by figures in authority, and will resist following a rule given by a figure in authority if they see this rule as unduly harmful or unfair. Moreover, they will cite these ethically relevant features to explain their resistance.[44] At the same age, children will spontaneously share their gains from a joint activity with a co-participant to redress an unfair distribution or unwarranted punishment by a figure in authority,[45] and will spontaneously attempt to stop third-party ethical transgressions.[46] Just as infants are to a considerable degree autonomous, experience-based causal learners[47] and learners of theory of mind,[48] capable of forming without explicit

instruction non-egocentric representations and evaluations of their causal, social, and epistemic environment, so infants appear to a considerable extent to be autonomous, experience-based ethical learners, capable of forming without explicit instruction the kinds of non-egocentric representations and evaluations of situations, agents, actions, and outcomes upon which responsiveness to ethically relevant features is based—as manifest, for example, in the social skills and maturity mentioned earlier. There is a dark side to such socially oriented learning: as infants gain in sophistication about social relations, they become more oriented toward what they find familiar or, somewhat later, toward people they perceive as members of their own group. However, while debate persists on this question, such "own group" preference does not appear to be a "wired-in" response as such and does not prevent infants or adults from being capable of an extraordinary degree of spontaneous cooperation and collaboration with unrelated individuals, especially in comparison with our nearest animal relatives.[49] And social learning in settings involving shared activities and goals can help counteract implicit bias.[50]

But what if we look beyond the developmental setting? What evidence do we have of the kinds of capacities that could underlie the ethical judgments of adults? Here we will briefly consider two kinds of evidence, from neuroimaging studies of ethical judgment and from informal classroom sampling of "ethical intuitions."

The question of the neural basis for ethical judgments has generated a large volume of research, the general trend of which has only fairly recently become clear. Initially, partly under the influence of innatist notions of a "moral module," it was thought that there might be some region or regions of the brain specialized for ethical judgment. By contrast, the approach to ethical development sketched here would predict that the neural substrate of ethical judgment would involve regions or networks subserving general-purpose learning and judgment concerning a range of causal and theory-of-mind–related questions about situations, actions, outcomes, and agents. Recently, metastudies of experimental reports of neural imaging during ethical judgment have come to the conclusion that ethical judgment relies heavily upon just such a neural network of regions, the *default network*.

The "default mode" of brain functioning is one of two primary modes of brain activity, alternating with the more focused "attentional mode."[51] Each mode corresponds to higher levels of coordinated activation in a relatively stable, interconnected set of brain regions. What are someof the functions of default network processing? They include, most importantly, episodic and semantic memory, scene construction and the imaginative simulation of possible futures, counterfactual reasoning, inferring the mental states of others, self-referential processing, and ethical judgment.[52] In other words, the primary network subserving

ethical judgment has the features that would be predicted by a model of ethical development as continuous with these other forms of cognition and evaluation.

There are of course many complexities and pitfalls in any appeal to neuroimaging evidence, and we can distinguish multiple kinds of ethical judgment, such as active versus passive, self-referring versus other-referring, and intuitive versus deliberative.[53] It is therefore still much too early to have any definitive picture of the neural basis of ethical thought and feeling. But neuroimaging using a variety of techniques has thus far been largely consistent with the idea that ethical cognition is supported by domain-general processing and essentially continuous with other ways in which we size up situations and actions and make evaluations and choices.[54]

More broadly, neuroimaging and connectivity research have increasing put in question the kind of "affective versus cognitive" division of mental processing found in many "dual-process" models of ethical cognition.[55] There are indeed forms of processing located in regions of the brain associated with affect that interact early and quickly with sensory input, before higher-order declarative reasoning has begun to operate, but these are also systems that subserve probabilistic learning, spatial mapping, evaluative comparison, and other core elements of "cognition."[56] Increasingly, a picture is emerging of cognition as widely distributed in the brain, and the age-old idea of the mind as pitting "reason" against "emotion" may be an artifact of our limited insight into the ways in which our minds actually operate. "Affect," as psychologists understand it, is not simply a matter of aroused emotion but is a capacity of the brain to synthesize multiple streams of information and evaluation in a manner that can orient or reorient a suite of mental processes—attention, perception, memory, inference, motivation, action-readiness—in a coordinated way to address actual or anticipated challenges.[57] If we are asking how an artificial system might make intelligent decisions responsive to ethically relevant features, we may wish to emulate the functional characteristics of this design,[58] which is inherited from our animal ancestors and highly conserved evolutionarily.

"Ethical intuitions" have also been subject to extensive research in recent decades. "Intuition" here does not designate a specific kind of mental process as such, but rather an assessment—whether of a particular scenario, a type of action, or a general principle—that is often relatively fast and effortless yet that typically feels compelling even though we have little insight into the process by which we arrived at it and may be unable to articulate a satisfactory rationale for it.

A principal focus of discussions of ethical intuition in recent decades, and of discussions of ethics and artificial intelligence as well, has been the "Trolley Problem," a puzzling pattern of ethical intuitions reliably evoked by a series of scenarios involving runaway trolleys. Trolley problems have sometimes been

called the *Drosophila* of ethical inquiry—a shared, heavily studied "test bed" for hypotheses about ethical judgment. It is, moreover, a nice irony that trolley problems, long castigated by critics as hopelessly artificial, turn out to have such direct analogues in one of the most important actual applications of artificial intelligence to life-affecting decision-making to date: self-driving vehicles. Let us begin, then, to look at trolley problems to see whether we can discover anything of relevance to our discussion of the nature and origin of human responsiveness to ethically relevant features. I believe we can.

To make my argument, I will be drawing upon in-class, confidential sampling of the intuitive ethical judgments of undergraduates in large ethics lectures I have taught at the University of Michigan over a number of years. During lecture, students are able to respond rapidly and confidentially to questions I pose by using individual wireless keypads (iClickers) that transmit their responses to a receiver at the front of the room. I am then able to display the overall patterns of response on a screen for students to see. These are hardly controlled experiments, and so they must be considered suggestive only.[59] But their informality also has advantages, in that it enables me to push a bit beyond the usual tightly constrained diet of standard examples in the trolley literature, and perhaps to probe a bit beneath the surface of my students' responses.

I needn't here rehearse the particulars of the most familiar trolley problems, which we will call "Switch" and "Footbridge."[60] In their responses to Switch and Footbridge, my students typically exhibit the same pattern of response that has been found repeatedly in the literature. In Switch, a strong majority (typically about 80%) say that one *should* push a lever to switch the runaway trolley to a sidetrack, saving five workers down the main track but killing one worker on the sidetrack. And in Footbridge, a strong majority (typically about 75%) says that one *should not* push a large man off a footbridge to stop the runaway trolley to save five workers. Despite a certain abstract similarity of the two scenarios—in both, an intervention taken to prevent the deaths of the five workers on the main track brings about the death of one other individual who is not initially at risk—the asymmetry in intuitive judgment has proven remarkably robust. Even moral philosophers who have considered the problem for years and who themselves judge that one *should* push the man off the footbridge tend to admit that this scenario does not cease to trouble them. Since the trolley problems first emerged in the 1970s, dramatic changes have occurred in people's views about interracial marriage, women's roles, gay marriage, premarital sex, smoking marijuana, and more. Yet the trolley problem asymmetry remains pretty much undiminished. Further, the asymmetry has been found cross-culturally, doesn't manifest gender differences, and appears both in vivid virtual-reality simulations and in simple, undramatic, verbal posing of the dilemmas.[61]

The problem continues to fascinate because there has been no analysis of the asymmetry that has received wide acceptance, despite many attempts.[62] One promising early explanation—roughly, that in Footbridge one is deliberately using the worker killed as a "mere means," whereas the worker dies in Switch as an "unintended side effect"—has lost adherents owing to a case called "Loop": Suppose the switch could send the trolley down a sidetrack that loops back to the main track; however, this will stop the trolley from hitting the five workers because a single, large worker is currently on the sidetrack, and the trolley, hitting him, will stop before rejoining the main track. Should you switch the trolley, killing one in order to save five? Here, according to the standard interpretation, the single worker on the sidetrack is being used as a "means" in essentially the same way as "Footbridge," since his being struck by the trolley is not an unintended side effect but essential to saving the five. Despite this, a strong majority of my students (typically about 80%), and of most populations sampled, say one *should* push the lever to send the trolley onto the looping sidetrack.[63]

At this point "dual-process" psychologists entered the fray, arguing that the difference between Switch and Loop, on the one hand, and Footbridge, on the other, is attributable not to a matter of ethical principle but to a rapid, strong, automatic, affectively charged, negative System 1 (or, more recently, "model-free") reaction to the thought of using direct muscular force to kill the man in Footbridge. This rapid "push button" response does not occur in cases like Switch and Loop, where the victim is less proximate and one's effect upon the victim less direct, so that the System 1 response is relatively weak, and a slower and more deliberative System 2 (or "model-based") response can come into play, favoring a calculation of minimizing harm.[64] This dual-process account affords an explanation of the asymmetry, but not one that provides much by way of ethical justification. Hence, some have argued on this basis that we should discount the normative significance of the Footbridge verdicts.[65]

But now consider "Beckon": As before, the runaway trolley will strike and kill five workers if not stopped. You are at some distance from the track, with no access to a switch, but you see a large man standing on the other side of the track, facing in your direction but unable to see the trolley approaching. If you conspicuously beckon to the man, encouraging him vigorously to come in your direction, he will step onto the track and immediately be struck and killed by the trolley, stopping it before it hits the five workers. In classroom sampling, I have regularly found that 60% to 70% of students say that one *should not* beckon to the man. This despite the fact that the death he suffers happens at a distance and involves no direct exertion of my own muscular force upon him.

Is intentionally gesturing in a way that lures someone to his death the problem? Consider now "Wave": You are standing down the track from the five workers, who are looking in your direction and do not see the trolley approaching them

from behind. If you wave vigorously to the side, encouraging them to step in that direction, the five workers will step off the track and be saved. However, another worker who is looking your way and who is initially standing *alongside* the track will also see your waving gesture and step in the same direction. This will place him on the track, where he will be struck from behind and killed. Here some 70% to 90% of my students will say that one *should* wave to the five workers, saving them but killing the one man lured thereby onto the track. What, then, could explain *this* asymmetry, which is as pronounced as the original Switch versus Footbridge asymmetry?

Suppose for argument that the earlier account of ethical judgment—as involving general capacities for modeling, simulating, and evaluating situations, actions, agents, and outcomes—was accepted. This would suggest that we should look for a complex competence in understanding the social landscape and its possibilities underlying all these trolley problems. How might we find out? When I ask my students whether learning that their roommate has been in a Switch-like trolley problem and has pulled the lever to send the trolley down the sidetrack would increase, decrease, or not affect the *trust* they have in their roommate, the majority response typically is "no change in trust," while "increase trust" and "decrease trust" each receives a smaller number of votes. When a similar question about trust is asked about a roommate who took action in Loop, student answers are essentially the same. But when asked about learning that a roommate has pushed a large man off a footbridge in a Footbridge scenario, the strong majority response (typically 70% to 80%) is "decrease trust," with a much smaller number indicating "no change" and virtually no one indicating "increase trust." In fact, in a typical sample, a much smaller number indicate "increase trust" in the Footbridge case (around 5%) than had originally judged that one *should* push the man (around 25-30%).

So now, what about Wave and Beckon? Here the response in Wave is essentially indistinguishable from that in Switch and Loop, while the response in Beckon is essentially indistinguishable from that in Footbridge. As in Footbridge, a smaller number indicate "increase trust" (about 5%) than initially judged one should take the action in question (about 35%). This pattern of trust judgments has been found each year I've sampled my students, reliably grouping Switch, Loop, and Wave into one category with regard to trust, and Footbridge and Beckon into another.

Perhaps, then, my students' intuitive responses to individual trolley scenarios involves not simply thinking about the *act* involved but thinking "What kind of person would perform this act?" and perhaps "Would I?" Personality tests have been given to subjects about to be given trolley problems, and those giving a "push" response in Footbridge as a group, in comparison to the group giving a "don't push" response, scored on average higher on psychopathy scales

and higher in indifference to harm or to ethical violations generally, while they scored lower on perspective-taking and altruism.[66] It would seem that my students' trustworthiness judgments may be tracking something real about "the kind of person who would perform this act."

But why would this consideration show up in an intuitive sense of what one should do in a given scenario? Suppose, as *virtue theorists* such as Aristotle[67] and Hume[68] have argued, our primary access to our ethical understanding is not via highly general principles or judgments of particular acts, but via our general sense of the tendencies of certain kinds of traits of character or motivational structures. Looked at from a modeling perspective, one might think one gains greater predictive and explanatory purchase in ethical thought if one assesses those around one in terms of their general dispositions to act or their trustworthiness. To gain an idea of how to act in a given situation, then, it may be more reliable to ask whether someone who manifested skills and traits of character we'd ethically admire would perform the act.

To further examine this interpretation, I ask my students what emotions they would expect to feel, had they intervened in a trolley problem to save the five and afterward decided to approach the family of the single victim they had killed. In the case of Switch, Loop, and Wave, the predominant response is to anticipate feeling regret and *guilt*, with some expectation that the family might understand. In the cases of Footbridge and Beckon, the predominant response is to anticipate feeling regret and *shame*, with little or no expectation that the family would understand. Anticipated shame, as opposed to anticipated guilt, suggests that they think others would also think that performing the interventions in Footbridge and Beckon would be a sign of defective character.

My hypothesis is that, when making an ethical assessment, my students (and the rest of us) rely upon acquired, general, abstract causal-evaluative models of situations and agents to simulate possible actions and likely outcomes or reactions. The simulations can be quite complex: *How would it feel to perform this action? Could I actually see myself doing it? What kind of person would perform it? What would others think, and could I face them?* But this kind of real-time simulation and evaluation of possibilities, and associated feelings and reactions on the part of others is exactly the kind of *prospective* processing the human default system appears to be engaged in systematically, off and on throughout the day, as we navigate the physical and social environment.[69]

This picture of intuitive ethical judgment also fits the recent proposal that prospection is a fundamental organizing principle of the human brain.[70] And it echoes the idea that prediction is of the essence in learning and intelligence, whether animal or machine. Relatedly, several recent studies of ethical judgments[71] have found that a model of the hypothetical agent and choice seems to mediate the "intuitive" judgment of the action.[72] If some elements of people's

acquired causal-evaluative models of situations and agents are based upon extensive, ordinary experience of a kind most people could be expected to have, whatever their social or cultural identity, this would explain how some patterns of intuitive ethical judgment, such as the trolley asymmetries, could be found very widely and remain stable across a number of social or cultural changes. It would also help explain why the source of such patterns might be difficult to introspect, and why the patterns might nonetheless remain confident even though they cannot be fit to a priori ethical principles.[73]

This brings us to the "realistic Trolley Problem" that has been much discussed in connection with self-driving cars. I have polled my students about two possible <situation, action> rules that might be "programmed into" self-driving cars with one passenger aboard: (1) they might be programmed to swerve to avoid five individuals in a cross-walk, even in cases where this would result in the death of one other individual, not now at risk, on a side walkway; (2) they might be programmed to swerve to avoid five individuals in a cross-walk, even in cases where this would result in colliding with a concrete wall, killing the occupant in the car. When I first posed these questions to students several years ago, a strong majority agreed with programming (1) but disagreed with programming (2). This initially seemed to replicate the kind of asymmetry found in Switch versus Footbridge or Wave versus Beckon. However, over the intervening years the percentage approving programming (1) has remained consistently high, at 70% to 80%, but the percentage approving programming (2) has climbed from 35% to 65%.

Why has the original asymmetry not been robust? One potential explanation: these are cases in which the *agent* and questions about the *character* of the agent have been removed from the situation. Initially students might have been tempted to assimilate self-driving cars to personified agents, but as the problem of regulating self-driving cars became more familiar over time, students became more likely to think of the problem in terms of *general rules*, and from that standpoint, there seems to be no reason to assign special weight to the car's occupant over pedestrians. Interestingly, the initial asymmetry actually went away over the course of the term, as discussion of the case proceeded. By contrast, the initial trolley asymmetry, and that between Wave and Beckon, tend to persist from one end of the term to the other, despite extensive discussion.

1.6. Artificial Ethical Psychology

The burden of the argument thus far is that we should understand the human capacity to identify and respond to ethically relevant considerations as an integral part of the competencies and knowledge we acquire that underwrite

human general intelligence and capacity for open-ended problem-solving. The reasoning has drawn heavily on evidence from human psychology, but many elements of the argument do not turn on details specific to *Homo sapiens*. For example, the ways in which default, defeasible trust or cooperation can make possible positive-sum results in learning, language, and social interactions depend upon very generic dynamics of agents and groups of agents.

Perhaps our model of how to develop *machine ethics* should not be based upon the idea of "programming in" principles or designing machines to "align themselves" with the preferences or values of the humans they encounter. Neither of these seems to be the way in which humans acquire ethical competence. It is not primarily "inculcated" into children by explicit adult teaching; indeed in many societies there is relatively little direct instruction of children. And, as we have seen, children display greater autonomy than simply aligning themselves with the preferences or values of the adults around them. Humans are hardly ideals of ethical competence, but if we wish to develop machines at least as trustworthy with life-affecting decision-making as ordinary humans, perhaps we should look to models of the development of machine ethics that more closely approximate human ethical learning.

In truth, we do not know what the principles would be for "programming in" ethics as anything like an operational system. There is continuing disagreement over the fundamental principles of ethics, and even supposing this were not so, there is sufficient distance between fundamental principles and actual applications (*What constitutes a harm in a given instance? How to assess the relative magnitude of harms and benefits? When have the conditions of a promise been sufficiently undermined that it no longer binds?*) that a large quantity of ethical understanding is needed in order to apply them—understanding that seems to come only with extensive individual and shared experience and is not contained within the principles themselves. Even if we consider a disaggregated system of less fundamental ethical rules, the actual situations we encounter are too varied, and the kinds of considerations that need to be taken into account too diverse, to allow these rules to be more than rules of thumb. Perhaps the ethical theory closest to common sense in contemporary Western society is W. D. Ross's system of prima facie duties,[74] but it is a fundamental feature of Ross's account that these duties can come into conflict and that there are no strict rules for determining which duties are weightier in a given case. Instead, Ross argued, we must have recourse to *intuition*.

It is a striking fact about ethical judgment that it has such a strong intuitive element, even in the assessment of ethical theories. What might a model of our ethical competence look like that would make sense of this idea of intuition? While some philosophers have thought of intuition as something like direct rational perception of self-evident truths, a long tradition in philosophy holds that

intuition is more like common sense—a large body of understanding, relying heavily upon experience and social discussion, without the structure of a deductive system, yet carrying enough structured information to make nuanced judgment possible. Fortunately, recent research in artificial intelligence is beginning to give us an idea of what such a large body of intuitive understanding or common sense might look like, how it might be acquired through experience, and how it might support abstract generalizations as well as nuanced judgments of novel cases, despite lacking an overall deductive, rule-like structure.[75] Rather than rely upon preprogrammed feature detectors or policies, programs that have been successful in tasks such as image identification, natural-language processing and translation, game playing, and motor control have been able to acquire high levels of competence via processes of learning based upon autonomous development of complex, generative representations of large bodies of data.

Here we have speculated that, in the case of humans, a relatively modest set of priors—for example, curiosity and default, defeasible reliance and trust—could combine with basic faculties and ample learning capacity to promote the acquisition of predictive and generative representations of the physical and social world, for example, intuitive physics, intuitive psychology, or communicative competence. We speculated further that these same capacities can subserve epistemic and ethical evaluation, and gave some evidence from neuroscience and ordinary ethical judgment to support this speculation. Thinking about machine ethics may need to undergo the same kind of "learning revolution" that thinking about machine learning and expertise, and thinking in developmental psychology, have undergone in recent years.

How might artificial ethical learning proceed? Here I have no expertise. Fortunately, however, this question is already being examined under other descriptions. For example, machines that are learning to carry on effective natural-language conversations—for example, to provide customer service that satisfactorily identifies and addresses problems, or to provide companionship and reliable health monitoring for an elderly person living at home, or to help students learn by identifying their strengths and weaknesses and drawing upon their abilities and motivations—are acquiring skills in understanding people and their needs and aims, and learning what it is like to work together with people to achieve mutually desirable outcomes. As artificial systems are increasingly deployed in our lives, the *social* dimension of their existence—their ability to work together with humans and together with one another—will become increasingly important and a fundamental part of the enlargement of their intelligence. We should be asking how systems, equipped with such priors as curiosity and default, defeasible trust or cooperativeness, might come to be themselves complex social agents, subject to demands that require them to figure out how to achieve solutions by working with others, sharing out tasks and assigning

responsibility in ways that can achieve positive sums and help sustain further cooperation—*learning* fundamental elements of ethics and knitting them into their global knowledge and competence.

Just as infants observe countless hours of adult behavior seeking to predict what will happen next, machines can observe countless hours of human and machine behavior seeking to predict, first, the next instant, then, the next second, then the next minute, and so on. They can learn to read the goals and beliefs of those around them, learning, as "mature" children do, such skills as recognizing the needs of those around them, even those who depart from the norm, or standing up for their own interests while according weight to the interests of others, or entering into novel relations and helping others to do so without conflict, and so on.[76] Adversarial training might pit machines observing human interactions and making predictions, or telling credible stories, against humans doing the same, asking the discriminator to determine whether the source is artificial or human. Machines can be apprentices or partners in complex tasks, learning social as well as technical skills. Self-driving cars, for example, can learn how to manage the elaborate interactions involved in a crowded parking lot at holiday time, or merging into bridge traffic at rush hour, in ways that achieve a successful mix of forcefulness and deference, reading the intentions of other drivers or autonomous vehicles in order to find workable solutions. Like humans, machines can use their internal models to create non-egocentric as well as egocentric representations and evaluations of situations, actions, outcomes, and policies. Like humans, machines can use these models to maintain some degree of autonomy in evaluation and action. We already know that machines should not be built so that they will pursue whatever goal they are given unquestioningly. Intelligent machines, like intelligent animals, should operate with modulated uncertainty rather than absolute certainty, and should be able to use their own resources, and draw upon others as a resource, for criticism and self-criticism.

Human learning is most impressive when it leverages the ability to form communicative and cooperative relations with others that extend our problem-solving capacity far beyond whatever we individuals could accomplish on our own. Artificial learning likewise can reach its fullest development socially and cooperatively, drawing upon an expanding network of perspectives and experience. The threat of an emergent "superintelligence" or, much more proximately, of artificial intelligence working in the service of those who'd rather dominate and exploit than work together and share, can only be met by developing a sufficiently robust community of cooperating human and artificial intelligences that takes advantage of the fact that a society capable of joint effort and sharing is in the long run likely to know more, and adapt more readily and with greater foresight, than a society based upon subordinating the interests of the many to the

interests of the few, and the suppression of alternative points of view. We can only speculate, but from the perspective of learning, it would seem that humans are more valuable as cooperation partners than as peons or fodder. Indeed superintelligent machines themselves should be able to see this, especially if we've had the sense to enable them to grow up socially, as our partners in learning.[77]

Notes

1. The use of the terms 'ethical' and 'moral' often gives rise to puzzlement over the relation between the two. In empirical psychology, where emphasis tends to be placed upon the person and personality, the terms 'moral' and 'morality' are most often used—e.g., "moral development" and "moral judgment". By contrast, in many areas of normative application the terms 'ethical'and 'ethics' are most often used, e.g., "medical ethics" and "ethics and artificial intelligence". In philosophy, and in this paper, these terms are used almost interchangeably, though 'ethical' often includes *prudence* as well as morality proper. E. Turiel, *The Culture of Morality: Social Development, Context, and Conflict* (Cambridge: Cambridge University Press, 2002); J. G. Smetana et al., "Developmental Changes and Differences in Young Children's Moral Judgments," *Child Development* 83 (2012): 683–96.
2. R. Frank, *Passions within Reason: The Strategic Role of the Emotions.* (New York: W.W. Norton, 1989); J. Decety and A. N. Meltzoff, "Empathy, Imitation, and the Social Brain," in *Empathy: Philosophical and Psychological Perspectives*, ed. A. Copland and P. Goldie (New York: Oxford University Press, 2011), 58–81.
3. N. Bostrom, *Superintelligence: Paths, Dangers, Strategies* (Oxford: Oxford University Press, 2014).
4. Oderbank, S.G., J. Nitschke, A. Mokros, E. Habermeyer, and O. Wilhelm, "Psychopathic Men: Deficits in General Mental Ability, Not Emotion Perception," *Journal of Abnormal Psychology* 127 (2018): 294-304; Kavish, N., C. Bailey, C. Sharp, and A. Venta, "On the Relation between General Intelligence and Psychopathic Traits: An Examination of Inpatient Adolescents," *Child Psychiatry and Human Development* 49 (2018): 341-51.
5. R. J. R. Blair, "The Amygdala and Ventromedial Prefrontal Cortex in Morality and Psychopathy," *Trends in Cognitive Sciences* 11 (2007): 387–92; R. J. R. Blair, "The Emergence of Psychopathy: Implications for the Neurophysiological Approach to Developmental Disorders," *Cognition* 101 (2006): 414–42.
6. Monagan, C., H. Bizumic, and M. Sellbom, "Nomological Network of Two-Dimensional Machiavellianism," *Personality and Individual Differences* 130 (2018): 161-72.
7. Bereczkei, T., P. Papp, P. Kinces, B. Bodrogi, G. Perlaki, G. Orsi, and A. Deak, "The Neural Basis of the Machiavellians' Decision Making in Fair and Unfair Situations," *Brain and Cognition* 98 (2015): 53-64.

8. Henrich, J., R. Boyd, S. Bowles, C. Camerer, E. Fehr, and H. Gintis. *The Foundations of Human Sociality: Economic Experiments and Ethnography Evidence from Fifteen Small-Scale Societies*. Oxford: Oxford University Press, 2004.
9. M. Hoffman, *Empathy and Moral Development: Implications for Caring and Justice* (Cambridge: Cambridge University Press, 2001); Decety and Meltzoff, "Empathy, Imitation, and the Social Brain."
10. A. Van den Oord, Y. Li, and O. Vinyals, "Representation Learning with Contrastive Predictive Coding," *arXiv*, last revised January 22, 2019, arXiv:1807.03748v1; J. Devlin, M.-W. Chang, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," *arXiv*, last revised May 24, 2019, arXiv:1810.04805v1,https://arxiv.org/abs/1810.04805.
11. S. Pinker, *The Language Instinct* (New York: Morrow, 1994); E. S. Spelke and K. D. Kinzler, "Innateness, Learning, and Rationality," *Child Development Perspectives* 3 (2009): 96–98.
12. G. K. Pullum and B. C. Scholz, "Empirical Assessment of Stimulus Poverty Arguments," *Linguistic Review* 19 (2002): 9–50.
13. J. B. Tenenbaum et al., "How to Grow a Mind: Statistics, Structure, and Abstraction," *Science* 331 (2011): 1279–85.
14. Van den Oord, Li, and Vinyals, "Representation Learning with Contrastive Predictive Coding"; Devlin, Chang, and Toutanova, "BERT."
15. R. N. Aslin, J. R. Saffran, and E. L. Newport, "Computation of Conditional Probability Statistics by 8-Month-Old Infants," *Psychological Science* 9 (1998): 321–24; C. Kidd, S. T. Piantadosi, and R. N. Aslin, "The Goldilocks Effect: Human Infants Allocate Attention to Sequences That Are Neither Too Simple nor Too Complex," *PLOS-One* 7 (2012): e36399.
16. T. Behne et al., "Unwilling versus Unable: Infants' Understanding of Intentional Action," *Development Psychology* 41 (2005): 328–37.
17. M. Carpenter, J. Call, and M. Tomasello, "Twelve- and 18-Month-Olds Copy Actions in Terms of Goals," *Developmental Science* 8 (2005): F13–F20.
18. J. Moll and M. Tomasello, "12- and 18-Month-Olds Follow Gaze to Spaces behind Barriers," *Developmental Science* 7 (2004): F1–F9.
19. F. Warneken, and M. Tomasello, "Altruistic Helping in Human Infants and Young Chimpanzees," *Science* 311 (2006): 1301–3.
20. J. Halberda, "The Development of a Word-Learning Strategy," *Cognition* 87 (2003): B23–B34.
21. Tenenbaum, J. B., C. Kemp, T. L. Griffiths, and N. D. Goodman. "How to Grow a Mind: Statistics, Structure, and Abstraction." *Science* 331 (2011): 1279–85; H. M. Wellman, *Making Minds: How Theory of Mind Develops* (Oxford: Oxford University Press, 2014); Goodman, N. D., J. B. Tenenbaum, J. Feldman, and T. L. Griffiths. "A Rational Analysis of Rule-Based Concept Learning." *Cognitive Science* 32 (2008): 108–54. Goodman, N.D., M.C. Frank, T.L. Griffiths, J.B. Tenenbaum, P.W. Battaglia, and J.B. Hamrick, "Relevant and Robust: A Response to Marcus and Davis," *Psychological Science* 26 (2015): 539-41.
22. R. C. Conant, and W. R. Ashby, "Every Good Regulator of a System Must Be a Model of That System," *International Journal of Systems Science* 1 (1970): 89–97.

23. E. Todorov and Z. Ghahramani, "Unsupervised Learning of Sensory-Motor Primitives," *Proceedings of the 25th Annual International Conference of the IEEE EMBS* (2003): 1750–53; E. Todorov, "Optimality Principles in Sensorimotor Control," *Nature Neuroscience* 7 (2004): 907–15; O.-S. Kwon and D. C. Knill, "The Brain Uses Adaptive Internal Models of Scene Statistics for Sensorimotor Estimation and Planning," *PNAS* 110, no. 11 (2013): E1064–E1073, https://doi/10.1073/pnas.1214869110.
24. A. Gopnik and H. Wellman, "Reconstructing Constructivism: Causal Models, Bayesian Learning, and the Theory Theory," *Psychological Bulletin* 128 (2012): 1085–108.
25. Wellman, *Making Minds*.
26. A. N. Meltzoff, "'Like Me': A Foundation for Social Cognition," *Developmental Science* 10 (2007): 126–34; J. N. Saby, A. N. Meltzoff, and P. J. Marshall, "Infant's Somatotopic Neural Responses to Seeing Human Actions: I've Got You under My Skin," *PLOS ONE* 8, no. 10 (2013): e77905, https://doi:10.1371/journal.pone.0077905.
27. H. M. Wellman and D. Liu, "Scaling of Theory-of-Mind Tasks," *Child Development* 75 (2004): 523–41.
28. E. I. Moser, E. Kropff, and M.-B. Moser, "Place Cells, Grid Cells, and the Brain's Spatial Representation System," *Annual Review of Neuroscience* 31 (2008): 69–89.
29. A. Johnson, M. A. A. van der Meer, and A. D. Redish, "Integrating Hippocampus and Striatum in Decision-Making," *Current Opinion in Neurobiology* 17 (2007): 692–97; A. S. Gupta et al., "Hippocampal Replay Is Not a Simple Function of Experience," *Neuron* 65 (2010): 695–705; A. D. Redish, "Vicarious Trial and Error," *Nature Reviews: Neuroscience* 17 (2016): 147–59.
30. A. N. Meltzoff et al., "Foundations for a New Science of Learning," *Science* 325 (2009): 284–88.
31. M. A. Koenig, V. Tiberius, and K. Hamlin, "Children's Judgments of Epistemic and Moral Agents: From Situations to Intentions," unpublished manuscript.
32. Ibid.; though on the role of native predispositions versus learned preferences in early filial responses, see E. Di Giorgio et al., "Filial Responses as Predisposed and Learned Preferences: Early Attachment in Chicks and Babies," *Behavioural and Brain Research* 325 (2017): 90–104.
33. H. Gweon and L. Schulz, "16-Month-Olds Rationally Infer Causes of Failed Actions," *Science* 332 (2011): 1524.
34. Warneken and Tomasello, "Altruistic Helping in Human Infants and Young Chimpanzees"; R. Roth-Hanania, M. Davidov, and C. Zhan-Waxler, "Empathy Development from 8 to 16 Months: Early Signs of Concern for Others," *Infant Behavior and Development* 34 (2011): 447–58.
35. Warneken and Tomasello, "Altruistic Helping in Human Infants and Young Chimpanzees."
36. Thomas Hobbes, *Leviathan* (1651), ed. C. B. MacPherson (London: Penguin, 1968).
37. R. Axelrod and D. Dion, "The Further Evolution of Cooperation," *Science* 242 (1988): 1385–90.
38. C. Boehm, *Moral Origins: The Evolution of Virtue, Altruism, and Shame* (New York: Basic Books,. 2012).

39. J. K. Rilling et al., "A Neural Basis for Social Cooperation," *Neuron* 36 (2002): 395–406; D. G. Rand, J. D. Greene, and M. A. Nowak, "Spontaneous Giving and Calculated Greed," *Nature* 489 (2012): 427–30; M. Crockett et al., "Harm to Others Outweighs Harm to Self in Moral Decision Making," *PNAS* 111 (2014): 17320–25; O. FedlmanHall et al., "Empathic Concern Drives Costly Altruism," *NeuroImage* 105 (2015): 347–56.
40. Warneken and Tomasello, "Altruistic Helping in Human Infants and Young Chimpanzees."
41. S. Benhabib and F. Dallmayr, eds., *The Communicative Ethics Controversy* (Cambridge, MA: MIT Press, 1990).
42. C. Peterson et al., "Peer Social Skills and Theory of Mind in Children with Autism, Deafness, or Typical Development," *Developmental Psychology* 52 (2016): 46–57.
43. Koenig, Tiberius, and Hamlin, "Children's Judgments of Epistemic and Moral Agents," Unpublished manuscript (2019)
44. Turiel, *The Culture of Morality*; Smetana et al., "Developmental Changes and Differences in Young Children's Moral Judgments."
45. N. Chernyak and D. M. Sobel, "'But He Didn't Mean to Do It': Preschoolers Correct Punishments Imposed on Accidental Transgressors," *Cognitive Development* 39 (2016): 13–20.
46. Vaish, A., M. Missana, and M. Tomasello. "Three-Year-Old Children Intervene in Third-Party Moral Transgressions." *British Journal of Developmental Psychology* 29 (2011): 124–30.
47. D. M. Sobel and N. Z. Kirkham, "Bayes' Nets and Babies: Infants' Developing Statistical Reasoning and Their Representation of Causal Knowledge," *Developmental Science* 10 (2007): 298–306; D. M. Sobel and N. Z. Kirkham, "Blickets and Babies: The Development of Causal Reasoning in Toddlers and Infants," *Developmental Psychology* 42 (2006): 1103–15.
48. Wellman, *Making Minds*.
49. Y. Bar-Haim et al., "Nature and Nurture in Own-Race Face Processing," *Psychological Science* 17 (2006): 159–63; F. Warneken and M. Tomasello, "The Roots of Human Altruism," *British Journal of Psychology* 100 (2009): 455–71; H. Over, "The Influence of Group Membership on Young Children's Prosocial Behavior," *Current Opinion in Psychology* 20 (2018): 17–20.
50. T. F. Pettigrew and L. R. Tropp, "A Meta-analytic Test of Intergroup Contact Theory," *Journal of Personality and Social Psychology* 90 (2006): 751–83; N. Dasgupta and L. M. Rivera, "When Social Context Matters: The Influence of Long-Term Contact and Short-Term Exposure to Admired Outgroup Members on Implicit Attitudes and Behavioral Intentions," *Social Cognition* 26 (2008): 112–23.
51. R. L. Buckner, J. R. Andrews-Hanna, and D. L. Schacter, "The Brain's Default Network: Anatomy, Function, and Relevance to Disease," *New York Academy of Sciences* 1124 (2008): 1–38.
52. Ibid.; G. Sevinc, and R. N. Spreng, "Contextual and Perceptual Brain Processes Underlying Moral Cognition: A Quantitative Meta-analysis of Moral Reasoning and

Moral Emotions," *PLOS ONE* 9, no. 2 (2014): e87427, https://doi:10.1371/journal/pone.0087427.

53. R. L. E. P. Reniers et al., "Moral Decision-Making, ToM, Empathy, and the Default Mode Network," *Biological Psychiatry* 90 (2012): 202–10; W. Chiong et al., "The Salience Network Causally Influences Default Mode Network Activity during Moral Reasoning," *Brain* 136 (2013): 1929–41; B. Garrigan, A. L. R. Adlam, and P. E. Langton, "Neural Correlates of Moral Decision-Making: A Systematic Review and Meta-analysis of Moral Evaluations and Response Decision Judgments," *Brain and Cognition* 108 (2016): 88–97, corrigendum, *Brain and Cognition* 111 (2016): 104–6.
54. A. Rangell, C. Camerer, and P. R. Montague, "A Framework for Studying the Neurobiology of Value-Based Decision-Making," *Nature Reviews: Neuroscience* 9 (2008): 545–56; T. E. J. Behrens et al., "Associative Learning of Social Value," *Nature* 456 (2008): 245–50; A. Shenhav and J. D. Greene, "Moral Judgments Recruit Domain-General Valuation Mechanisms to Integrate Representations of Probability and Magnitude," *Neuron* 67 (2010): 667–77; F. A. Cushman, and L. Young, "Patterns of Moral Judgment Derive from Nonmoral Psychological Representations," *Cognitive Science* 35 (2011): 1052–75.
55. J. D. Greene et al., "An fMRI Investigation of Emotional Engagement in Moral Judgment," *Science* 293 (2001): 2015–18; J. Greene and J. Haidt, "How (and Where) Does Moral Judgment Work?," *Trends in Cognitive Sciences* 6 (2002): 517–23; J. D. Greene et al., "Pushing Moral Buttons: The Interaction between Personal Force and Intention in Moral Judgment," *Cognition* 111 (2009): 364–71.
56. L. Pessoa, "On the Relationship between Emotion and Cognition," *Nature Reviews Neuroscience* 9 (2008): 148–58.
57. S. R. Quartz, "Reason, Emotion, and Decision-Making: Risk and Reward Computation with Feeling," *Trends in Cognitive Sciences* 13 (2007): 209–15; A. D. Craig, "How Do You Feel—Now? The Anterior Insula and Human Awareness," *Nature Reviews Neuroscience* 10 (2009): 59–70; R. M. Nesse and P. E. Ellsworth, "Emotion, Evolution, and Emotional Disorders," *American Psychologist* 64 (2009): 129–39.
58. Cf. Marvin Minsky, *The Emotion Machine: Commonsense Thinking, Artificial Intelligence, and the Future of the Human Mind* (New York: Simon and Schuster, 2006).
59. For some studies supporting the reliability of such electronic in-class sampling, see J. R. Stowell and J. M. Nelson, "Benefits of Electronic Audience Response Systems on Student Participation, Learning, and Emotion," *Teaching of Psychology* 34 (2007): 253–58; G. E. Kennedy and Q. I. Cutts, "The Association between Students' Use of an Electronic Voting System and Their Learning Outcomes," *Journal of Computer Assisted Learning* 21 (2005): 260–68.
60. J. J. Thomson, "Killing, Letting Die, and the Trolley Problem," *Monist* 59 (1976): 205–17.
61. N. Gold, A. M. Colman, and B. D. Pulford, "Cultural Differences in Responses to Real-Life and Hypothetical Trolley Problems," *Judgment and Decision Making* 9 (2014): 65–76; C. D. Navarette et al., "Virtual Morality: Emotion and Action in a Simulated 'Trolley Problem,'" *Emotion* 12 (2012): 364–70.

62. For an especially sophisticated discussion, see F. Kamm, *Intricate Ethics: Rights, Responsibilities, and Permissible Harm* (New York: Oxford University Press, 2007).
63. Though on the potential influence of order effects on Loop verdicts, see S. M. Liao et al., "The Brain Uses Adaptive Internal Models of Scene Statistics for Sensorimotor Estimation and Planning," *PNAS* 110, no. 11 (2013): E1064–E1073, https://doi/10.1073/pnas.1214869110.
64. Greene et al., "An fMRI Investigation of Emotional Engagement in Moral Judgment"; F. A. Cushman, "Action, Outcome, and Value in a Dual-System Framework for Morality," *Personality and Social Psychology Review* 17 (2013): 273–92.
65. Greene, J.D., F.A. Cushman, L.E. Steward, K. Lowenberg, L.E. Nystrom, and J.D. Cohen, "Pushing Moral Buttons: The Interaction between Personal Force and Intention in Moral Judgments," *Cognition* 111 (2009): 364-71.
66. Bartells, D.M. and D.A. Pizarro, "The Mismeasure of Morals: Antisocial Personality Traits Predict Utilitarian Responses to Moral Dilemmas," *Cognition* 121 (2012): 490-503; Kahane, G., J. A. C. Everett, B. D. Earp, M. Farias, and J. Savulescu. "'Utilitarian' Judgments in Sacrificial Moral Dilemmas Do Not Reflect Impartial Concern for the Greater Good." *Cognition* 134 (2015): 193–209.; Y. Gao and S. Tang, "Psychopathic Personality and Utilitarian Moral Judgment in College Students," *Journal of Criminal Justice* 41 (2013): 342–49; P. Conway and B. Gawronski, "Deontological and Utilitarian Inclinations in Moral Decision Making: A Process Dissociation Approach," *Journal of Personality and Social Psychology* 104 (2013): 216–35; E. Gleichgerrcht and L. Young, "Low Levels of Empathic Concern Predict Utilitarian Moral Judgment," *PLOS-One* 8 (2013): e60418. But see also for qualifications P. Conway, J Goldstein-Greenwood, D. Polacek, and J.D. Green, "Sacrificial Utilitarian Judgments Do Reflect Concern for the Greater Good: Clarification via Process Dissociation and the Judgments of Philosophers," *Cognition* 179 (2018): 241–65..
67. Aristotle, *Nicomachean Ethics* (350–340 BCE), trans. T. Irwin, 2nd ed. (Indianapolis, IN: Hackett, 1999).
68. David Hume, *An Enquiry concerning the Principles of Morals* (1751), ed. T. L. Beauchamp (Oxford: Oxford University Press, 1998); David Hume, *A Treatise of Human Nature* (1738), ed. L. A. Selby-Bigge and P. H. Nidditch (Oxford: Oxford University Press, 1978).
69. Buckner, Andrews-Hanna, and Schacter, "The Brain's Default Network."
70. M. E. P. Seligman et al., "Navigating into the Future or Driven by the Past?," *Perspectives in Psychological Science* 8 (2013): 119–41.
71. Including a study of the "Knobe Effect": C. S. Sripada, "Mental State Attributions and the Side-Effect Effect," *Journal of Experimental Social Psychology* 48 (2012): 232–38.
72. E. L. Uhlmann, L. Zhu, and D. Tannenbaum, "When It Takes a Bad Person to Do the Right Thing," *Cognition* 126 (2013): 326–34.
73. For further discussion, see Peter Railton, "The Affective Dog and Its Rational Tale: Intuition and Attunement," *Ethics* 124 (2014): 813–59; Peter Railton, "Moral Learning: Conceptual Foundations and Normative Significance," *Cognition* 167 (2016): 172–90.
74. W. D. Ross, *The Right and the Good* (Oxford: Oxford University Press, 1930).

75. For an example in the confined world of games, see D. Silver et al., "A General Reinforcement Learning Algorithm Masters Chess, Shogi, and Go through Self-Play," *Science* 362 (2018): 1140–44.
76. Cf. Peterson et al., "Peer Social Skills and Theory of Mind in Children with Autism, Deafness, or Typical Development."
77. The author would like to thank participants in the NYU Conference on Ethics and Artificial Intelligence (October 2016) for very helpful discussions, including especially Ned Block, Paul Boghossian, Nick Bostrum, David Chalmers, Vasant Dhar, Yann LeCun, S. Matthew Liao, Stuart Russell, Wendell Wallach, and Stephen Wolfram. I would also like to thank my colleagues Sarah Buss, Ben Kuipers, and Chandra Sripada for insightful conversations and sustaining encouragement.

References

Aristotle. *Nicomachean Ethics*. 350–340 BCE. Translated by T. Irwin. 2nd ed. Indianapolis, IN: Hackett, 1999.

Aslin, R. N., J. R. Saffran, and E. L. Newport. "Computation of Conditional Probability Statistics by 8-Month-Old Infants." *Psychological Science* 9 (1998): 321–24.

Axelrod, R., and D. Dion. "The Further Evolution of Cooperation." *Science* 242 (1988): 1385–90.

Bar-Heim, Y., T. Ziv, D. Lamy, and R. M. Hodes. "Nature and Nurture in Own-Race Face Processing." *Psychological Science* 17 (2006): 159–63.

Behne, T., M. Carpenter, J. Call, and M. Tomasello. "Unwilling versus Unable: Infants' Understanding of Intentional Action." *Development Psychology* 41 (2005): 328–37.

Behrens, T. E. J., L. T. Hunt, M. W. Woolrich, M. F. S. Rushworth. "Associative Learning of Social Value." *Nature* 456 (2008): 245–50.

Benhabib, S., and F. Dallmayr, eds. *The Communicative Ethics Controversy.* Cambridge, MA: MIT Press, 1990.

Bereczkei, T., P. Papp, P. Kinces, B. Bodrogi, G. Perlaki, G. Orsi, and A. Deak, "The Neural Basis of the Machiavellians' Decision Making in Fair and Unfair Situations," *Brain and Cognition* 98 (2015): 53-64.

Blair, R. J. R. "The Amygdala and Ventromedial Prefrontal Cortex in Morality and Psychopathy." *Trends in Cognitive Sciences* 11 (2007): 387–92.

Blair, R. J. R. "The Emergence of Psychopathy: Implications for the Neurophysiological Approach to Developmental Disorders." *Cognition* 101 (2006): 414–42.

Boehm, C. *Moral Origins: The Evolution of Virtue, Altruism, and Shame.* New York: Basic Books,. 2012.

Bostrom, N. *Superintelligence: Paths, Dangers, Strategies*. Oxford: Oxford University Press, 2014.

Buckner, R. L., J. R. Andrews-Hanna, and D. L. Schacter. "The Brain's Default Network: Anatomy, Function, and Relevance to Disease." *New York Academy of Sciences* 1124 (2008): 1–38.

Carpenter, M., J. Call, and M. Tomasello. "Twelve- and 18-Month-Olds Copy Actions in Terms of Goals." *Developmental Science* 8 (2005): F13–F20.

Chernyak, N., and D. M. Sobel. "'But He Didn't Mean to Do It': Preschoolers Correct Punishments Imposed on Accidental Transgressors." *Cognitive Development* 39 (2016): 13–20.

Chiong, W., S. M. Wilson, M. D'Esposito, A.S. Kayser, S.N. Grossman, P. Poorzand, W.W. Seeley, B.L. Miller, and K.P. Rankin. "The Salience Network Causally Influences Default Mode Network Activity during Moral Reasoning." *Brain* 136 (2013): 1929–41.

Conant, R. C., and W. R. Ashby. "Every Good Regulator of a System Must Be a Model of That System." *International Journal of Systems Science* 1 (1970): 89–97.

Conway, P., and B. Gawronski. "Deontological and Utilitarian Inclinations in Moral Decision Making: A Process Dissociation Approach." *Journal of Personality and Social Psychology* 104 (2013): 216–35.

Conway, P., J. Goldstein-Greenwood, D. Polacek, and J. D. Greene. "Sacrificial Utilitarian Judgments Do Reflect Concern for the Greater Good: Clarification via Process Dissociation and the Judgments of Philosophers." *Cognition* 179 (2018): 241–65.

Craig, A. D. "How Do You Feel—Now? The Anterior Insula and Human Awareness." *Nature Reviews Neuroscience* 10 (2009): 59–70.

Crockett, M., Z. Kurth-Nelson, J. Z. Siegel, P. Dayan, and R. J. Dayan. "Harm to Others Outweighs Harm to Self in Moral Decision Making." *PNAS* 111 (2014): 17320–25.

Cushman, F. A. "Action, Outcome, and Value in a Dual-System Framework for Morality." *Personality and Social Psychology Review* 17 (2013): 273–92.

Cushman, F. A., and L. Young. "Patterns of Moral Judgment Derive from Nonmoral Psychological Representations." *Cognitive Science* 35 (2011): 1052–75.

Dasgupta, N., and L. M. Rivera. "When Social Context Matters: The Influence of Long-Term Contact and Short-Term Exposure to Admired Outgroup Members on Implicit Attitudes and Behavioral Intentions." *Social Cognition* 26 (2008): 112–23.

Decety, J., and A. N. Meltzoff. "Empathy, Imitation, and the Social Brain." In *Empathy: Philosophical and Psychological Perspectives*, edited by A. Copland and P. Goldie, 58–81. New York: Oxford University Press, 2011.

Devlin, J., M.-W. Chang, and K. Toutanova. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." *arXiv*, last revised May 24, 2019. arXiv:1810.04805v1. https://arxiv.org/abs/1810.04805.

Di Giorgio, E., J. L. Loveland, U. Mayer, O. Rosa-Salva, E. Versace, and G. Vallortigara. "Filial Responses as Predisposed and Learned Preferences: Early Attachment in Chicks and Babies." *Behavioural and Brain Research* 325 (2017): 90–104.

FeldmanHall, O., T. Dalgleish, D. Evans, and D. Mobbs. "Empathic Concern Drives Costly Altruism." *NeuroImage* 105 (2015): 347–56.

Frank, R. *Passions within Reason: The Strategic Role of the Emotions*. (New York: W.W. Norton, 1989).

Gao, Y., and S. Tang. "Psychopathic Personality and Utilitarian Moral Judgment in College Students." *Journal of Criminal Justice* 41 (2013): 342–49.

Garrigan, B., A. L. R. Adlam, and P. E. Langton. "Neural Correlates of Moral Decision-Making: A Systematic Review and Meta-analysis of Moral Evaluations and Response Decision Judgments." *Brain and Cognition* 108 (2016): 88–97; corrigendum, *Brain and Cognition* 111 (2016): 104–6.

Gleichgerrcht, E., and L. Young. "Low Levels of Empathic Concern Predict Utilitarian Moral Judgment." *PLOS-One* 8 (2013): e60418.

Gold, N., A. M. Colman, and B. D. Pulford. "Cultural Differences in Responses to Real-Life and Hypothetical Trolley Problems." *Judgment and Decision Making* 9 (2014): 65–76.

Goodman, N. D., J. B. Tenenbaum, J. Feldman, and T. L. Griffiths. "A Rational Analysis of Rule-Based Concept Learning." *Cognitive Science* 32 (2008): 108–54.

Goodman, N.D., M.C. Frank, T.L. Griffiths, J.B. Tenenbaum, P.W. Battaglia, and J.B. Hamrick, "Relevant and Robust: A Response to Marcus and Davis," *Psychological Science* 26 (2015): 539-541.

Gopnik, A., and H. Wellman. "Reconstructing Constructivism: Causal Models, Bayesian Learning, and the Theory Theory." *Psychological Bulletin* 128 (2012): 1085–108.

Greene, J., and J. Haidt. "How (and Where) Does Moral Judgment Work?" *Trends in Cognitive Sciences* 6 (2002): 517–23.

Greene, J. D., F. A. Cushman, L. E. Stewart, K. Lowenberg, L. E. Nystrom, and J. D. Cohen. "Pushing Moral Buttons: The Interaction between Personal Force and Intention in Moral Judgment." *Cognition* 111 (2009): 364–71.

Greene, J. D., R. B. Sommerville, L. E. Nystrom, J. M. Darley, and J. D. Cohen. "An fMRI Investigation of Emotional Engagement in Moral Judgment." *Science* 293 (2001): 2015–18.

Gupta, A. S., M. A. A. van der Meer, D. S. Touretzky, and A. D. Redish. "Hippocampal Replay Is Not a Simple Function of Experience." *Neuron* 65 (2010): 695–705.

Gweon, H., and L. Schulz. "16-Month-Olds Rationally Infer Causes of Failed Actions." *Science* 332 (2011): 1524.

Halberda, J. "The Development of a Word-Learning Strategy." *Cognition* 87 (2003): B23–B34.

Henrich, J., R. Boyd, S. Bowles, C. Camerer, E. Fehr, and H. Gintis. *The Foundations of Human Sociality: Economic Experiments and Ethnography Evidence from Fifteen Small-Scale Societies*. Oxford: Oxford University Press, 2004.

Hobbes, Thomas. *Leviathan*. 1651. Edited by C. B. MacPherson. London: Penguin, 1968.

Hoffman, M. *Empathy and Moral Development: Implications for Caring and Justice*. Cambridge: Cambridge University Press, 2001.

Hume, David. *An Enquiry concerning the Principles of Morals*. 1751. Edited by T. L. Beauchamp. Oxford: Oxford University Press, 1998.

Hume, David. *A Treatise of Human Nature*. 1738. Edited by L. A. Selby-Bigge and P. H. Nidditch. Oxford: Oxford University Press, 1978.

Johnson, A., M. A. A. van der Meer, and A. D. Redish. "Integrating Hippocampus and Striatum in Decision-Making." *Current Opinion in Neurobiology* 17 (2007): 692–97.

Kahane, G., J. A. C. Everett, B. D. Earp, M. Farias, and J. Savulescu. "'Utilitarian' Judgments in Sacrificial Moral Dilemmas Do Not Reflect Impartial Concern for the Greater Good." *Cognition* 134 (2015): 193–209.

Kavish, N., C. Bailey, C. Sharp, and A. Venta, "On the Relation between General Intelligence and Psychopathic Traits: An Examination of Inpatient Adolescents," *Child Psychiatry and Human Development* 49 (2018): 341-51.

Kamm, F. *Intricate Ethics: Rights, Responsibilities, and Permissible Harm*. New York: Oxford University Press, 2007.

Kennedy, G. E., and Q. I. Cutts. "The Association between Students' Use of an Electronic Voting System and Their Learning Outcomes." *Journal of Computer Assisted Learning* 21 (2005): 260–68.

Kidd, C., S. T. Piantadosi, and R. N. Aslin. "The Goldilocks Effect: Human Infants Allocate Attention to Sequences That Are Neither Too Simple nor Too Complex." *PLOS-One* 7 (2012): e36399.

Koenig, M. A., V. Tiberius, and K. Hamlin. "Children's Judgments of Epistemic and Moral Agents: From Situations to Intentions." Unpublished manuscript (2019).

Kwon, O.-S., and D. C. Knill. "The Brain Uses Adaptive Internal Models of Scene Statistics for Sensorimotor Estimation and Planning." *PNAS* 110, no. 11 (2013): E1064–E1073. https://doi/10.1073/pnas.1214869110.

Liao, S. M., A. Wiegmann, J. Alexander, and G. Vong. "Putting the Trolley in Order: Experimental Philosophy and the Loop Case." *Philosophical Psychology* 25 (2012): 661–71.

Meltzoff, A. N. "'Like Me': A Foundation for Social Cognition." *Developmental Science* 10 (2007): 126–34.

Meltzoff, A. N., P. K. Kuhl, J. Movellan, and T. J. Sejnowski. "Foundations for a New Science of Learning." *Science* 325 (2009): 284–88.

Minsky, Marvin. *The Emotion Machine: Commonsense Thinking, Artificial Intelligence, and the Future of the Human Mind*. New York: Simon and Schuster, 2006.

Moll, J., and M. Tomasello. "12- and 18-Month-Olds Follow Gaze to Spaces behind Barriers." *Developmental Science* 7 (2004): F1–F9.

Monagan, C., H. Bizumic, and M. Sellbom, "Nomological Network of Two-Dimensional Machiavellianism," *Personality and Individual Differences* 130 (2018): 161-72.

Moser, E. I., E. Kropff, and M.-B. Moser. "Place Cells, Grid Cells, and the Brain's Spatial Representation System." *Annual Review of Neuroscience* 31 (2008): 69–89.

Navarrette, C. D., M. M. McDonald, M. L. Mott, and B. Asher. "Virtual Morality: Emotion and Action in a Simulated 'Trolley Problem.'" *Emotion* 12 (2012): 364–70.

Nesse, R. M., and P. E. Ellsworth. "Emotion, Evolution, and Emotional Disorders." *American Psychologist* 64 (2009): 129–39.

Oderbank, S.G., J. Nitschke, A. Mokros, E. Habermeyer, and O. Wilhelm, "Psychopathic Men: Deficits in General Mental Ability, Not Emotion Perception," *Journal of Abnormal Psychology* 127 (2018): 294-304.

Over, H. "The Influence of Group Membership on Young Children's Prosocial Behavior." *Current Opinion in Psychology* 20 (2018): 17–20.

Pessoa, L. "On the Relationship between Emotion and Cognition." *Nature Reviews Neuroscience* 9 (2008): 148–58.

Peterson, C., V. Slaughter, C. Moore, and H. M. Wellman. "Peer Social Skills and Theory of Mind in Children with Autism, Deafness, or Typical Development." *Developmental Psychology* 52 (2016): 46–57.

Pettigrew, T. F., and L. R. Tropp. "A Meta-analytic Test of Intergroup Contact Theory." *Journal of Personality and Social Psychology* 90 (2006): 751–83.

Pinker, S. *The Language Instinct*. New York: Morrow, 1994.

Pullum, G. K., and B. C. Scholz. "Empirical Assessment of Stimulus Poverty Arguments." *Linguistic Review* 19 (2002): 9–50.

Quartz, S. R. "Reason, Emotion, and Decision-Making: Risk and Reward Computation with Feeling." *Trends in Cognitive Sciences* 13 (2007): 209–15.

Railton, Peter. "The Affective Dog and Its Rational Tale: Intuition and Attunement." *Ethics* 124 (2014): 813–59.

Railton, Peter. "Moral Learning: Conceptual Foundations and Normative Significance." *Cognition* 167 (2016): 172–90.

Rand, D. G., J. D. Greene, and M. A. Nowak. "Spontaneous Giving and Calculated Greed." *Nature* 489 (2012): 427–30.

Rangell, A., C. Camerer, and P. R. Montague. "A Framework for Studying the Neurobiology of Value-Based Decision-Making." *Nature Reviews: Neuroscience* 9 (2008): 545–56.

Reniers, R. L. E. P., R. Corcoran, B. A. Vollm, A. Mashru, R. Howard, and P. F. Liddle. "Moral Decision-Making, ToM, Empathy, and the Default Mode Network." *Biological Psychiatry* 90 (2012): 202–10.

Redish, A. D. "Vicarious Trial and Error." *Nature Reviews: Neuroscience* 17 (2016): 147–59.

Rilling, J. K., D. A. Gutman, T. R. Zeh, G. Pagnoni, G. S. Berns, and C. D. Kilts. "A Neural Basis for Social Cooperation." *Neuron* 36 (2002): 395–406.

Ross, W. D. *The Right and the Good*. Oxford: Oxford University Press, 1930.

Roth-Hanania, R., M. Davidov, and C. Zhan-Waxler. "Empathy Development from 8 to 16 Months: Early Signs of Concern for Others." *Infant Behavior and Development* 34 (2011): 447–58.

Saby, J. N., A. N. Meltzoff, and P. J. Marshall. "Infant's Somatotopic Neural Responses to Seeing Human Actions: I've Got You under My Skin." *PLOS ONE* 8, no. 10 (2013): e77905. https://doi:10.1371/journal.pone.0077905.

Seligman, M. E. P., P. Railton, R. Baumeister, and C. S. Sripada. "Navigating into the Future or Driven by the Past?" *Perspectives in Psychological Science* 8 (2013): 119–41.

Sevinc, G., and R. N. Spreng. "Contextual and Perceptual Brain Processes Underlying Moral Cognition: A Quantitative Meta-analysis of Moral Reasoning and Moral Emotions." *PLOS ONE* 9, no. 2 (2014): e87427. https://doi:10.1371/journal/pone.0087427.

Shenhav, A., and J. D. Greene. "Moral Judgments Recruit Domain-General Valuation Mechanisms to Integrate Representations of Probability and Magnitude." *Neuron* 67 (2010): 667–77.

Silver, D., T. Hubert, J. Schrittweiser, I. Antonoglou, M. Lai, A. Guez, M. Lanctot, L. Sifre, D. Kumaran, T. Graepel, T. Lillicrap, K. Simonyan, and D. Hassabis. "A General Reinforcement Learning Algorithm Masters Chess, Shogi, and Go through Self-Play." *Science* 362 (2018): 1140–44.

Smetana, J. G., W. M. Rote, M. Jambon, M. Tasopoulos-Chan, M. Villalobos, and J. Comer. "Developmental Changes and Differences in Young Children's Moral Judgments." *Child Development* 83 (2012): 683–96.

Sobel, D. M., and N. Z. Kirkham. "Bayes' Nets and Babies: Infants' Developing Statistical Reasoning and Their Representation of Causal Knowledge." *Developmental Science* 10 (2007): 298–306.

Sobel, D. M., and N. Z. Kirkham. "Blickets and Babies: The Development of Causal Reasoning in Toddlers and Infants." *Developmental Psychology* 42 (2006): 1103–15.

Spelke, E. S., and K. D. Kinzler. "Innateness, Learning, and Rationality." *Child Development Perspectives* 3 (2009): 96–98.

Sripada, C. S. "Mental State Attributions and the Side-Effect Effect." *Journal of Experimental Social Psychology* 48 (2012): 232–38.

Stowell, J. R., and J. M. Nelson. "Benefits of Electronic Audience Response Systems on Student Participation, Learning, and Emotion." *Teaching of Psychology* 34 (2007): 253–58.

Tenenbaum, J. B., C. Kemp, T. L. Griffiths, and N. D. Goodman. "How to Grow a Mind: Statistics, Structure, and Abstraction." *Science* 331 (2011): 1279–85.

Thomson, J. J. "Killing, Letting Die, and the Trolley Problem." *Monist* 59 (1976): 205–17.

Todorov, E. "Optimality Principles in Sensorimotor Control." *Nature Neuroscience* 7 (2004): 907–15.

Todorov, E., and Z. Ghahramani. "Unsupervised Learning of Sensory-Motor Primitives." *Proceedings of the 25th Annual International Conference of the IEEE EMBS* (2003): 1750–53.

Turiel, E. *The Culture of Morality: Social Development, Context, and Conflict*. Cambridge: Cambridge University Press, 2002.

Uhlmann, E. L., L. Zhu, and D. Tannenbaum. "When It Takes a Bad Person to Do the Right Thing." *Cognition* 126 (2013): 326–34.

Vaish, A., M. Missana, and M. Tomasello. "Three-Year-Old Children Intervene in Third-Party Moral Transgressions." *British Journal of Developmental Psychology* 29 (2011): 124–30.

Van den Oord, A., Y. Li, and O. Vinyals. "Representation Learning with Contrastive Predictive Coding." *arXiv*, last revised January 22, 2019. arXiv:1807.03748v1.

Warneken, F., and M. Tomasello. "Altruistic Helping in Human Infants and Young Chimpanzees." *Science* 311 (2006): 1301–3.

Warneken, F., and M. Tomasello. "The Roots of Human Altruism." *British Journal of Psychology* 100 (2009): 455–71.

Wellman, H. M. *Making Minds: How Theory of Mind Develops*. Oxford: Oxford University Press, 2014.

Wellman, H. M., and D. Liu. "Scaling of Theory-of-Mind Tasks." *Child Development* 75 (2004): 523–41.