Available online at www.sciencedirect.com
SCIENCE DIRECT
Cognitive Systems Research 7 (2006) 246–258
Cognitive Systems RESEARCH
www.elsevier.com/locate/cogsys

The economic and evolutionary basis of selves

Action editors: Luca Tummolini and Cristiano Castelfranchi

Don Ross [a,b,*]

[a] University of Alabama at Birmingham, 900 13th Street South, 4th Floor, Humanities Building, Birmingham, AL 35294-1260, USA
[b] University of Cape Town, School of Economics, Leslie Social Science Building, Private bag, Rondebosch 7701, South Africa

Received 23 March 2005; accepted 7 November 2005
Available online 28 February 2006

Abstract

This paper aims to reconcile radical anti-individualism about people, according to which people are dynamic products of social dynamics, with neoclassical economic formalism and standard evolutionary game theory. The point of doing so is to face empirical facts coming from the cognitive and behavioral sciences, without throwing away any more of our well established modeling technology than we have to. The paper develops a high-level framework for modeling *game determination*, the process by which people strategically interact (play games) to determine which ranges of subsequent games will be played by their future selves that will have been sculpted from the preference refinements resulting from the earlier games.

Keywords: Individual agency; Behavioral economics; Evolutionary game theory; Modeling; Selves; Social dynamics

1. Introduction

This paper forms part of a larger inquiry, whose main product to date is a book (Ross, 2005, 2006) on the integration of mainstream microeconomic theory with the other cognitive and behavioral sciences. To fix the context for the paper, I must first sketch some of the book's conclusions and their basis. The core contention is that new empirical and conceptual insights flowing out of the cognitive–behavioral sciences, including behavioral economics (e.g., Camerer, Loewenstein, & Rabin, 2003), should not stampede us into rejecting the formal framework of post-war neoclassical analysis, as has lately been urged by many authors (see especially Hodgson, 2001; Lawson, 1997; Sugden, 2001).[1] This is not because neoclassical theory gives a more promising account of human behavior than these authors think, or because it incorporates sound ontological principles for thinking about people. Rather, it is because neoclassical theory, properly understood, is not directly about any specific kind of behavior, and rests upon no ontological commitments more definite than the idea that agents can be analytically distinguished from one another. Radical methodological criticism of neoclassical theory generally rests on reading into it strong positive theses about individualism, and about human capacities for optimization and implementation of procedural rationality. I argue in the book that these readings confuse messages with messengers: from the fact that many theor*ists* have

* Tel.: +1 205 396 9071.
E-mail address: dross@commerce.uct.ac.za.

[1] Economists who promote evolutionary and behavioral–institutional economics, such as Gintis (2000) and Bowles (2004), often use a good deal of anti-neoclassical *rhetoric*, but, for reasons that will become clear shortly, aren't really foes of neoclassicism in the sense I intend here.

1389-0417/$ - see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.cogsys.2005.11.003

exhibited these commitments, it does not follow that the theory itself incorporates them.[2] If we are scrupulous about letting the mathematics do the talking, we can recover an adequate – indeed, very useful – analytical and conceptual framework from neoclassicism that strengthens, rather than introduces false assumptions into, behavioral inquiry, and is fully compatible with the empirical picture of people we are getting from the behavioral sciences.

In this paper my concern will just be with individualism. It is of course true that all of the early neoclassical theorists (e.g., Walras, Jevons, Fisher) were individualists – indeed, they took for granted a quite extreme form of early modern atomism, about people and everything else. But then, so did most of their contemporaries in every branch of inquiry, including Karl Marx (Elster, 1985) (but not their predecessor Adam Smith, as many are now coming to appreciate (Rothschild, 2001; Sugden, 2000)). Many postwar economists have been *normative* individualists, but this is a quite different thesis from the descriptive individualism – and its denial – that interests me here.

In the postwar formalism of neoclassicism there is no thesis at all about who or what agents are; an agent is simply anything whose behavior is well modeled within the constraints of a small set of consistency axioms. How long any given agent, so defined, lasts through a process is also a question on which the mathematics forces no stake. So individual humans can become new agents whenever their preferences change. Agents need not be internally simple – as people are not – so they can in principle be firms or households or whole countries or any other sort of unit that acts teleologically. I argue in the book (and in Ross, 2002) that *baseline* or *prototypical* cases of economic agency should indeed be simple – insects are good prototypes – but not for atomistic or reductionist reasons. The motivation, instead, is that such agents don't raise complications due to apparent preference reversals over their biographies; an entire biological bug *does* map relatively neatly onto representation as a single agent. Attention to properties of these baseline agents helps us fix state variables for use in more difficult extensions of the formalism to non-prototypical – more complex – agents like people and communities of people. Decades of experimental research by behavioral economists shows that individual biological instances of *H. sapiens* briefly instantiate particular economic agents only episodically, and thus resemble countries more than they do stable microeconomic units like bugs. Finally, I also argue in the book that non-prototypical agents are not mereological compositions of prototypical agents, thus hopefully blocking any temptation to read my suggestions about prototypical agency atomistically.

Debates over these issues should not be encouraged to degenerate into semantic exercises. If someone wants to reserve the adjective 'neoclassical' for a certain sort of ideological position, and say that any framework that doesn't encourage that ideological spin is then anti-neoclassical, I would prefer not to quarrel over application of the label. What I want to concentrate on here are some *substantive* alternatives in modeling that are obscured when core neoclassical formalism is read as if individualism were built into it. This encourages inferences that begin from the premise that (formal) neoclassical economic agents must be equivalent to individual human selves, work through a second premise that individual human willpower is not the main causal engine of human behavioral patterns, and get the conclusion that selves are epiphenomena so far as economic causation and explanation are concerned.

Variations on this inference are much in vogue, especially in the precincts of Santa Fe. The reasoning has been given explicit expression by Satz and Ferejohn (1994) and Sugden (2001). Their modus ponens has also been expressed as a modus tollens argument by Davis (2003), who first criticizes the idea that selves could be economic epiphenomena, and then moves backwards through the above premises, accepts the second one, and so concludes that the alleged neoclassical conception of agency should be rejected. I agree with Davis that selves are causally significant to human behavioral patterns, including economic ones, but since I reject the inference, I need not accept his *reductio* against the value of neoclassical theory. But I do accept the inference's second premise. Human behavioral patterns are mainly social and collective phenomena.

There are several complementary argument paths by which one can aim to keep this second premise of the inference, thus strongly denying descriptive individualism, while rejecting the inference's conclusion. In the book I explore historical and conceptual paths, and show how these can help us reinterpret experimental results from behavioral economics and cognitive psychology that have seemed to many to cast doubt on the significance of the self. Here, I will emphasize and extend one particular path, that of showing how we can model – and explain the stabilizing function of – selves in an ontological context that treats communities as logically prior for human behavioral explanation, while breaking none of the rules of the neoclassical formalism. The details of what is at stake in this enterprise of recovery will be indicated as I go along, but the main issue can be summarized upfront as follows. We should not suppose that we

[2] Defenders of 'bounded rationality' models of economic behavior, including many so-called 'behavioral' economists, disassociate themselves from neoclassicism on grounds that it is committed to optimization models. According to bounded rationality hypotheses, most economic behavior by people involves only 'satisficing' up to thresholds. However, as argued at length in Ross (2005, 2006), the rhetoric involved in presenting behavioral economics as *necessarily* a radical challenge to neoclassicism rests on conflation of two claims that are often similarly expressed but are logically distinct. Neoclassical formalism indeed demands models in which objective functions are maximized. However, this need not be interpreted as requiring that whole individual people act so as to maximize their stock of any material quantity. Indeed, what I take to be the canonical formulation of the core of neoclassical theory, Samuelson (1947), explicitly denies such an interpretation. Utility, as maximized in neoclassical models, is not a material quantity or an aggregate of material quantities, and models of material satisficers can always be expressed in terms of optimization functions. See Gintis (2004) for an instance of a leading behavioral economist stressing this point.

face a choice between an individualistic neoclassical economics and a non-individualistic, heterodox economics that eliminates selves as loci of causal significance. Instead, we can use a refined form of neoclassical analysis to explain how selves evolved to stabilize relationships of economic exchange.

2. Individualistic and non-individualistic models

Social and economic theory in the individualist tradition has tended to take stable 'selves' – used here to mean approximately 'individual human personalities' – as given, and then understand socialization as describable by some function that aggregates or otherwise composes these selves. Individualism may be thought of as coming in stronger or weaker versions depending on the nature of this function. Where the function is conceived to be linear, as in many neoclassical models, selves preserve their properties intact under socialization. The view that neoclassical models *must* restrict themselves to such functions – a view endorsed in their philosophical moments by a number of economists, though fewer than one might expect – has contributed to the interpretation of neoclassicism as necessarily individualistic. Models in the mainstream sociological and social–psychological traditions, on the other hand, often (but usually only implicitly) use non-linear composition functions. Marxism is less likely to be read as committed to individualism than is neoclassicism, despite the explicit individualism of its founder, because Marxist accounts typically invoke feedback from social phenomena that dynamically modifies the properties of selves. However, such models are still individualistic in what I will call the weak sense because they depict socialized identities as excretions of, or 'wrap-arounds' (Clark, 2002) to, pre-social selves. That is, they take individuals to be logically and ontogenetically prior to interactive networks. This distinction between strong and weak individualism, at least in economics, moves less furniture than is often supposed, because the models built around strong individualist assumptions can incorporate feedback with little strain. Individualistic economists and so-called 'analytic' Marxists thus have no serious difficulty talking to one another in a shared formal idiom (Roemer, 1981).

I will here take genuine *non-individualism* to be based on two increasingly widespread, and closely related, theses from cognitive–behavioral science. According to the first thesis (see Clark, 1997, 2004; Ross, 2005, 2006; Sterelny, 2004; Tomasello, 2001), human personalities – selves, that is – have been made phylogenetically possible and normatively central through the environmental manipulations achieved collectively by humans over their history, while *particular* people are ontogenetically created by cultural dynamics unfolding in this context. According to the second thesis, individual people are themselves systems governed by distributed–control dynamics (Ainslie, 2001; Clark, 1997; Dennett, 1991; Ross, 2005, 2006; Schelling, 1980), and so must for various explanatory and predictive purposes be modeled as bargaining communities. These theses together imply that adequate models of people – and not just of groups of people – will be social–dynamic models through and through.

Note that *interesting* individualism cannot just be the truism that biological individuals are importantly distinct from one another in all sorts of ways; so interesting anti-individualism is not the implausible denial of that truism. Furthermore, descriptive anti-individualism is made interesting *by* normative individualism: it is because we are *normatively* interested in human personalities, and not mainly in human organisms,[3] that the logical and ontogenetic basis of these personalities in social dynamics, and their implementation as bargaining communities, are such striking things to recognize.

The increasingly popular way of formally representing non-individualist models, both of communities of people and of the communities of interests that instantiate people, involves taking the communities in question to be dynamic systems. Such models are developed and promoted in a diverse range of literatures, including evolutionary game theoretic (EGT) approaches in economics (Gintis, 2000; Young, 1998), and applications of artificial life (AL) techniques to social theory (Kennedy & Eberhart, 2001). These are what I alluded to above as 'the precincts of Santa Fe'. My interest in defending the continuing usefulness of some neoclassical principles is inspired by foundational worries about an aspect of these approaches.[4]

Here is the worry. Whereas individualistic models incorporate pure fictions – strongly unified selves that are established prior to and independently of social dynamics – most current EGT and AL models commit an (implicit and usually unintended) oversimplification in the opposite direction. That is, in treating top-down dynamical influences as the sole sources of non-accidental causation, and then in addition modeling all dynamic phenomena as irreducibly statistical, they leave selves as, at best, mere cognitive book-keeping devices or, at worst, scientifically mysterious epiphenomena. A social theorist might argue that this needn't much trouble her, on grounds that models of social processes can reasonably abstract away from the influences of selves, the idiosyncracies of which might be washed out by the law of large numbers on her level of analysis. She need not *deny* that causally efficacious selves exist, but she can leave their analysis and study to personality psychologists working at microanalytic levels. I see three problems with this kind of response.

First, it ignores the possibility that selves might function as behavior-stabilization devices that in turn contribute (socially important) stabilizing properties to social dynamics. Second, the response fails to acknowledge that different

[3] For explanatory contrast: 'fundamentalist' moral positions on abortion and stem cell research that are politically influential in the United States fail to track this distinction.

[4] 'Aspect' needs emphasizing here; I am not encouraging skepticism about the value of these modeling approaches. I think that any young economist who does not learn evolutionary game theory thereby fails to acquire a piece of core professional technology.

social processes can produce alternative distributions of types of selves within communities,[5] and that this might in turn feed back to produce varying distributions of attractors in large-scale dynamics. Extant dynamic-systems models might thus fare poorly in predicting or explaining cross-cultural differences among people. Third, the response invites us to try to finesse rather than address some deep indeterminacies concerning the empirical applications of dynamic models. If we implicitly suppose that complex systems (like social communities) are dynamical systems 'all the way down', then it is unclear how, or whether, we might find general forms for writing down theories, as opposed to merely particular descriptions of histories that don't facilitate formal generalization. This seems to leave something missing from the ontological foundations of social theory. Individualist neoclassical models show us how we might formally restrict a concept of the self for use in economics: selves are associated with unchanging preference fields (e.g., Stigler & Becker, 1977). The anti-individualist argues that this won't work because it relies on utterly fictional objects. However, it is not obvious how to move beyond this purely negative point to state a positive alternative that can be given generalizable, non-metaphorical content. This worry does not merely represent philosophical nostalgia for theories written in a classical idiom. It expresses itself for practical purposes in EGT and AL models as instability of state variables from one model to another; in comparing these models, one gets the impression that decisions about what to regard as state variables are often driven as much by features of the software packages used for modeling as by any motivated theoretical principles. To the extent that this is so, it is hard to see how we could ever expect a convincing argument for regarding one model or family of models as empirically more persuasive than another.

My aim here is thus to present a sketch of the dynamics of self-formation under social pressure that (i) makes explicitly theorized selves emerge as causally and explanatorily significant, and endogenous, elements of social dynamics, without following individualism in taking them as primitive; and (ii) preserves a role for the neoclassical concept of utility, not as a representation of any empirical force or quantity, but simply as a formal organizing principle – analogous but not identical to 'fitness' in population ecology – that permits development of field theories to ontologically anchor dynamical accounts of societies. Neoclassical theory achieves a balance of the kind that has been of pervasive importance to the progress of science. On the one hand, it is representationally flexible enough to avoid pre-commitment to strongly restrictive ontological assumptions. As noted above, anything that pursues consistently maintained goals, even if just for a short interval, can be modeled as a neoclassical agent during that interval. And any set of non-contradictory goals at all can be represented by a utility function.[6] On the other hand, such constraints as *are* imposed by the mathematical properties of neoclassical systems are well understood. Thus we can compare the dynamical properties (relative sizes of basins of attraction, relative sensitivity to quantitative adjustments of parameters, etc.) of any two models constructed in this formalism with maximum clarity. This yields us the basic Popperian virtue of being able to *reject* models for sound empirical reasons.[7]

3. Selves

I derive my target concept of a 'self' from work by cognitive scientists and philosophers, specifically Bruner (1992, 2002), Dennett (1991, 2003) and Flanagan (2002). That is, I take selves to be narrated structures that enhance individuals' predictability, both to themselves and to others. As emphasized by this literature, selves in this respect resemble characters in novels and plays, in a number of quite specific ways. In particular, they facilitate increasing predictive leverage over time by acquiring richer structure as the narratives that produce them identify their dispositions in wider ranges of situations. On this account, individuals are not born with selves; furthermore, to the extent that the consistency constraints on self-narratives come from social pressures, particular narrative trajectories are not endogenous to individuals. As Dennett (1991) puts the point, selves have multiple authors, even if one author is most important in playing a role across all chapters while collaborators vary.

This philosophical account nicely captures the phenomenology and microstructure of selfhood. A personality is experienced to itself, and to others, as a relatively coherent story; to the extent that it is not, pathologies couched in terms of 'breakdown' tend to be diagnosed and – crucially – to trigger social sanctions. Stability is emphasized as an

[5] A newly flourishing area of empirical study is personality differences in non-human animals. Natural selection appears to strongly maintain such differences, at least in intelligent social species. Scientists aim to explain this (Bouchard & Loehlin, 2001).

[6] It might be objected that non-contradiction is in fact a poorly motivated restriction on the flexibility I just lauded. Isn't the whole point of emphasizing behavioral evidence in economics that we stop imposing ideals of rationality on the agents we model? Why should *this* aspect of the ideal of rationality continue to be privileged? I respond to this objection as follows. There are well-honed techniques in utility theory for handling limited, local contradictions. But to the extent that contradictions globally iterate across a system's behavioral patterns, the neoclassicists have been right to think that we should surrender our commitment to viewing the system in question as an agent. Resistance to this, where the systems in question are people, stems from an individualistic insistence that biological organisms must be prototypical agents. What else could it be based on? How could one have evidence that a system really is pursuing goals if the goals in question are taken to be globally contradictory? Those who take it as *axiomatic* that whole people are single agents across time thereby ignore the need for such evidence, which is why for them agency and consistency can seem to come apart.

[7] Some critics of mainstream economics (e.g., Rosenberg, 1992), argue that neoclassically inspired microeconomics doesn't much manifest this virtue. I think this is empirically false. The history of postwar economics is littered with models that were once thought to be promising accounts of phenomena and are now *known* to be false. Behavioral and experimental economics has played a leading role in such Popperian progress.

essential normative property of a self, though, just as with characters in novels, it is not literally to be maximized. (Totally consistent characters in novels are rejected as 'one-dimensional', and totally consistent people are sanctioned by being labeled 'boring' or, worse, 'obsessive'.) Instead, it is a background condition that makes some desirable extent of novelty from occasion to occasion meaningful and attractive. People evaluating and tinkering with their own personalities are usually acutely conscious of being monitored by others, and of being answerable to social norms and expectations, while doing so. I suggest that the close analogy between psychological and literary narrative is not just a fortuitous metaphor. As Elster (1999) has argued – and promoted into a methodological motif – literary narrative conventions are likely projections of natural psychological ones, and the creation of literary characters is modeled on the creation of selves. As has been understood since at least Aristotle's time, one can scientifically study some psychological dynamics by studying fiction. The narrative theory of the self helps to explain this otherwise odd fact.

It is not mysterious that natural selection in a social species like *H. sapiens* could give rise to narrative selves. Because of the complexity of their control systems – their brains *and* their networks of environmental pressures – people can't simply *assume* self-predictability; they have to act so as to *make* themselves predictable. They do this *so* they can play and resolve coordination games with others. (To be predictable to others, they must be predictable to themselves, and vice-versa.) Then all of this is compounded by the fact that nature doesn't neatly partition games the way analysts do in game theory texts. A person can't keep the various games she simultaneously plays with different people in encapsulated silos, so a move in a game G_i with the stranger will *also* represent moves in other games $G_{k,\ldots,n}$ with more familiar partners – because these partners are watching, and will draw information relevant to $G_{k,\ldots,n}$ from, what she does in G_i.

It is highly unlikely that the system of logical pressures set up by these dynamics is perfectly tractable by any finite information-processor in real time. People navigating in a web of social relationships face a continuous general equilibrium problem, and a non-parametric one at that. People probably do not *literally* solve such problems, that is, actually find optimal solutions to their sets of simultaneous games (except, sometimes, by luck). As discussed in Ross (2004, 2005, 2006), for example, the phenomenon of the 'mid-life crisis' picks out the pattern that manifests itself when people regret the formerly open possibilities their self-narratives have closed off, and so try to withdraw some but not all of their investment in their self; but then the bits of the portfolio turn out to be interdependent, so valued stock is unintentionally thrown away with what's deliberately discarded and personal disasters often result. However, most people achieve tolerable success as satisficers over the problem space. They do this at the cost of increasingly sacrificing flexibility in new game situations. This, happily, trades off against the fact that as their selves become more stable, they can send clearer signals to partners, thereby reducing the incidence of both miscoordination within assurance or coordination games, and of inadvertently stumbling into destructive Prisoner's Dilemma scenarios. This general fact itself helps to explain the prevailing stability of selves in a feedback relationship. It is sensible for people to avoid attempts at coordination with highly unstable selves. Given the massive interdependency among people, this incentivizes everyone to regulate the stability of those around them through dispensation of social rewards and punishments (Binmore, 1998, 2005; Ross, in press). Thus selves arise as stabilizing devices for social dynamics, and are in turn stabilized by those very dynamics.

However, from the anti-individualist perspective this notion of the self skates over a problem. That is, it seems as if the account requires us to treat individuals as primitive and then describe them as coming to be endowed with selves. In a neoclassical representation, socially sculpted selves will have to be assigned different utility functions from pre-social individuals, since self-sculpting must involve preference modification. Furthermore, if the subject's own (essential) participation in self-narration is a strategic response aimed at coordination with others, then an economic model must interpret selves as products of games played amongst sets of players that can't include that very self. We can't build models of these games unless we have players with well defined utility functions to start with. In that case, mustn't a traditional economic model of the dynamics underlying narrative selves be an instance of an individualistic model in the weak sense specified above, that is, one that takes pre-social individuals as logically and ontogenetically primary, though allowing non-linear composition functions in representing their interactions?

We can carve a path around this conceptual impasse by appeal to two sources of tools and ideas: (1) control theory from AI and neuroscience, and (2) behavioristic neoclassical consumer theory (as in Samuelson, 1947). Attention to (1) forces us to take seriously *some* limits on the sensitivity of behavior and agency to all the dynamical forces present in an environment. Complex systems can only manifest agency if they achieve stable integration of information in such a way as to shield them, up to a point, from dynamical perturbations. At the sub-personal level, nervous systems have access to information in more restrictive formats from those available to whole socialized people (Clark, 1990).[8] This enables neuroscientists to account for solutions to special bottleneck problems that arise in modeling the flow of

[8] The idea alluded to here will be familiar to philosophers of mind and to most cognitive scientists, but probably not to economists. To briefly explain it: communities of people use their shared public language to redescribe things, highlighting different aspects and relations of an event, process or object on different occasions while still (usually) keeping reference fixed. Coding at the neural level, not being public, can't be assigned this sort of flexible semantic interpretation. At the same time, people have no direct access to the coding system used by their brains. Thus we say that the two informational formats – neural coding and public language – are distinct, and usable by distinct systems.

information as it can be used by synaptic networks. (2) is useful because it encourages us to separate treatment of utility functions in extension from decision-theoretic conceptions of expected-utility maximization *processes* (that is, explicit 'in-board' calculations).

I will first explain the relevance of (2). Historically, it is the assumption that utility-maximization by people must consist in calculation of expected utility that has led neoclassical theorists to take for granted that if they're modeling anything empirically real, this must be direct behavior determination by whole people processing information formatted for personal-level use.[9] However, *nothing* in the neoclassical mathematics of the decade before Von Neumann – Morgenstern utility forces this interpretation. The importance of this point can be emphasized by attention to issues arising in the new 'neuroeconomics' literature (Glimcher, 2003; Montague & Berns, 2002) that studies individual neurons as economic agents. So far, this exciting work has not been generally careful about keeping personal-level informational content distinct from subpersonal-level content, and so encourages a slide back into an individualist conception in which people are taken to be mereologically composed out of functional modules that locally supervene on neuronal groups. This perspective is explicit in Glimcher (2003), who has had the most to say about the philosophical foundations of neuroeconomics.

Let us consider an example. Evidence reviewed by Montague and Berns (2002) suggests that firing rates of neurons in primate orbitofrontal cortex and ventral striatum encode a common currency by which primate brains can compare valuations of prospects over rewards in different modalities. The equation that describes the value the brain attaches to getting a particular predictor signal in a sequence of perceptions turns out to be a generalization of the Nobel-winning Black–Scholes model for pricing assets in financial markets with derivatives. Montague and Berns rightly express some enthusiasm about this fact, since the neural valuation equation and empirical tests of Black–Scholes respectively derive their data from wholly independent domains; the isomorphism between the equations is a discovery, not a construction. A natural explanation of the relationship might be that human investors are using their primate brains to estimate value, on which market prices are in turn based. Now, from this platform consider another striking empirical suggestion coming from neuroeconomic research with both monkeys and people. Montague and Berns report that when the predictor-valuation model is applied to subjects making risky investment decisions under uncertainty, subjects cluster strongly into two groups. One group plays optimally through runs of losses that could have been predicted to occur with positive probability at some point, while subjects in the other group abandon their portfolios too quickly for optimization of expected utility. The intriguing finding from the neuroeconomic research is that one can reliably predict which group a given subject will fall into by examining her brain under fMRI and determining whether neurons in her left nucleus accumbens respond to changes in the market data. 'Risk-takers' seem to be tracking predicted values explicitly with these neurons, while 'conservatives' may be falling back on more general heuristics that are biased against losses.

Montague and Berns themselves advance no philosophically loaded interpretations of these data. But it is easy to imagine such an interpretation, of just the sort that Glimcher (2003) encourages. Perhaps we should reduce explanations of people's risk-aversion levels to explanations of the risk-attitude dispositions of their brains. Imagine, for example, financial houses thinking that they should screen potential asset brokers under fMRI to make sure that they're not conservatives.

I think that sophistication about the philosophy of mind should discourage such interpretations. For a person, values of assets will be sensitive to ranges of parameters that are strongly controlled by social dynamics in which the person is embedded, but to which the brain won't be sensitive *at the same grain of analysis* as that at which it tracks frequency of perceptual cues; the person isn't identical to her brain because some counterfactuals relevant to generalizing about her behavior track regularities controlled by her social environment rather than (just) her nervous system. Of course, facts of the sort unearthed by neuroeconomists are relevant to our understanding of the *information* made available *to* the person *by* her brain. A broker who knows she has a conservative brain might have extra reason to rely more heavily on her computer model of asset price estimation than her colleagues whose brains do accurate tracking more directly. But conservative brains need not predict conservative selves.

Taking account of the way in which people are distinct from their brains is the point of my suggested appeal to neuroscientific control theory. In designing more sophisticated nervous systems over time – and thus encountering new risks of inefficiency due to bottlenecks – natural selection could not help itself to top-down control dynamics that arise when systems take the intentional stance (Dennett, 1987) towards themselves. Our pre-human ancestors could not assume this stance. Thus evolution had to solve the neural bottleneck problem *at the neural level*. On the other hand, accounts of selves as devices for integrating internal bargaining communities are often based partly on an argument to the effect that solving control problems in non-parametric environments is what selves evolved to do (Dennett, 1991; Ross, 2005, 2006; Sterelny, 2004). This precisely implies the distinction between *brain-level* individualism and *person-level* individualism, especially if one of the advantages people bring to the table by contrast with brains is faster response to the flexibility encoded in social learning. Brains bring compensating advantages of their own, as we should expect. As the discussion of asset valuation above suggests, their reduced plasticity relative to socially anchored selves can help maintain objectivity in circumstances where herd effects occur. It is just when we *don't* conflate maximization of utility by *brains*

[9] See previous note.

with goal achievement by *selves* that we have some hope of using data about the former as a source of theoretically independent constraints on processing models of the latter. Thus individualism about people could *impede* optimal use of neoclassical theory in a promising new domain of application – neuroeconomics – while a view that takes socially sculpted selves seriously as causal influences can focus our attention on what neoclassical theory is tailored for: describing the dynamics of information flow in markets.

So much for motivations. *How* might we try to take socially sculpted selves seriously in models that are both non-individualist and respectful of neoclassical restrictions on state variables?

4. Games, biological individuals, and people

I noted earlier that there is one sense of individualism that is truistic rather than false: there are biological individuals. Their important role in biological explanation is ensured by the fact that selection at the genetic level is non-Lamarckian, since although genetic selection is very far from all that is important in biological evolution (Keller, 1999), it obviously *is* important enough to make it essential for many purposes, both practical and theoretical, to keep track of individual phenomes. (See Buss, 1987 on the biology of individuals.) Let us then isolate the notion of biological individuality by reference to barriers on transmission of genetic information. (We won't need for present purposes to choose an explicit definition.) This gives us a basis both for treating species as kinds of individuals, following Hull (1976), and for taking organisms – but not people – to be individuals in the traditional sense.[10]

Now consider standard evolutionary games (Weibull, 1995). These will be games in which the expected distribution of strategies at any time t is a function of the expected fitnesses of strategies at $t-n, n \in \{\sum n, \ldots, \tau\}, t \leqslant \tau$.[11] Let G''_n denote such a game as played over n generations, where each generational cohort constitutes a 'round'. Let g''_{t+k} denote the $t+k$th round of G''_n, and model g''_{t+k} as a classical game. (Assume subgame perfection as the only refinement on Nash equilibrium in solving each $g''_n \subset G''_n$.) Let i, j denote biological individuals that are among the players of g''_{n+k} and indicate that i is a player of g''_{n+k} by writing $g''_{n+k(i)}$. In standard models of G''_n, no non-parametric problems need be solved *by* individual brains; all are solved only at the species level. This is a definitional, not an empirical, claim: it follows from understanding biological individuality in terms of barriers to genetic information flow. To see this: Define an 'agent' by reference to neoclassical formalism, that is, as a unit that has a utility function over the outcomes in which the payoffs of some set of games is specified. Distinguish an agent *i*'s 'agent-specific control system' as any nexus of causal influences on i that (i) is sensitive to values of strategic parameters in $g''_{n+k(i)}$ and (ii) exerts strategic influence on i without exerting influence on any player j of $g''_{n+k(i,j)}$ except via its influence on *i*'s strategy.[12] Then, unless selection is Lamarckian, no agent *i*'s agent-specific control system can introduce, at $n+k$, new information strategically relevant at round $g''_{n+k(i)}$ into an evolutionary game $G''_n \supset g''_{n+k(i)}$ in which i instantiates a strategy.[13] Otherwise, we would violate the assumption that no genetic information is transmitted between individuals except by ancestral descent.

Standard EGT models sometimes impose (at least implicitly) restrictions on the causal generation of organism behavior that are a bit weaker than the condition above. That is, they can allow for error terms (denoting what will appear as 'trembles' in a forecast of g''_{n+k+1} from $G''_{n\ldots n+k}$) inside which the causal influences of agent-specific control systems might figure. This would be necessary if one wanted to allow for possible mutations that affect dynamics by modifying cognitive dispositions. However, since such influences have to be non-autocorrelated with the original strategic dispositions of players in order to belong in error terms, this way of allowing for them can't introduce *systematic* roles for selves in evolutionary games. This is why some economists attracted to EGT models have suggested eliminativism about selves and properties of selves in discussions around the methodology of modeling social dynamics, as alluded to earlier (Sugden, 2001).

Restrictions of the sort I have just characterized are, in themselves, unobjectionable. All that they do is make explicit the idea that lineages, not organisms, are the proper players

[10] So, in terms of the earlier explanatory heuristic: a 'pro-life' fundamentalist is right to think that an abortion destroys the integrity of a biological individual, though it destroys the integrity of no individual person (which is what ought to matter morally).

[11] This formulation is not the standard textbook one. It deliberately abstracts away from some interesting issues about equilibrium computation in evolutionary games. n denotes the baseline point from which basins of attraction at t are calculated, and can be any distance into the past history of the lines of replicators interacting during the process. With respect to any model, we can ask about the extent to which its dynamics are path-dependent (that is, about the extent to which the law of large numbers will minimize the sensitivity of equilibria to low-probability events, as a function of n). To the extent that the model is relatively deterministic (not strongly path-dependent in its dynamics) we should get the same equilibria regardless of where we choose n. To the extent that we have strong path-dependency, we will get less fine-grained topologies of basins if we increase n while holding fixed the confidence level with which we want to calculate expected fitnesses. All equilibria as n moves closer to t will be refinements of earlier-n equilibria (i.e., consistent with those equilibria, but representing more information). In a particular model (e.g., any implemented simulation), n will be set by the modeler, so the game can be defined less generally than in the formulation here. These less general formulations are the usual ones found in textbooks.

[12] Condition (i) here restricts attention to causal influences that operate *by way of* *i*'s *agency* – telling us not to count, say, an asteroid that strikes only i as part of *i*'s agent-specific control system. Condition (ii) makes the control system specific to i. Note that the formulation is carefully neutral as between internalist and externalist interpretations of behavioral control, both in general and in particular circumstances.

[13] There is no restriction on biological individuals acting so as to introduce strategically relevant information into games going on amongst other agents.

of evolutionary games. Players of classically conceived rounds of these games are then either strictly deterministic products of the histories of prior rounds (as in the strict version of the restriction) or stochastic products of such rounds (as in the weaker version). The point I wish to stress by focusing on the restrictions is that in most applications of evolutionary economics it is implicitly supposed that games thus restricted provide sufficient explanations of all observed strategy dispositions. Note that this point applies not only to models in which strategies are taken to spread purely by vertical transmission (i.e., through inheritance), but also to models that are supposed to represent cultural evolution by assigning important roles to imitation. Imitation functions, of the sort studied by Young (1998), *amplify and stabilize* some effects of past strategy distributions on future ones, but they leave informational integration *by* organisms inside black boxes. The motivations for doing this are clear enough: to the extent that one tries to open these boxes, it seems one is no longer trying to get evolution to carry the explanatory burden, but is drifting back towards cognitive – and individualist – modeling.

Now we can more precisely characterize the challenge set in the preceding sections of the paper. Is there a way to pry open this black box in the modeling framework that lets agent-specific control systems (which might include selves, if these turn out to have any strategic function) exert causal influence on the dynamics of G''-type games while requiring that they emerge *endogenously* in these games? The question can be operationalized as follows: Can we build a well defined evolutionary game G''_n in which no properties of agent-specific control systems are relevant to computing equilibria at g''_{n+1} but some agents i, j arise through the dynamics of G''_n such that properties of the agent-specific control systems of i and j are relevant to computing equilibria in some $g''_{n+1+k(i,j)}$? (I specify both i and j here because I am interested not just in any old agent-specific control systems but in selves, and am assuming, for reasons discussed earlier, that selves can arise only as elements of reciprocal social relations. That is the very content of denying individualism.) Doing this would show us how to work selves into dynamic social games without having to make any such selves logically prior to the social dynamics.

Another distinction is now in order. Incorporating some conceptual suggestions of Sterelny's (2004), let us distinguish *social* dynamics, as dynamics that arise whenever biological individuals play games, from *cultural* dynamics, which presuppose the relevance of agent-specific control systems to evolutionary equilibria. (Sterelny argues that cultural accumulation requires the evolution of cognitive capacities that go beyond the mere capacity for imprinting on others as imitation targets, and permit decoupled representation[14] of goals and techniques based on others' behaviors as models of abstract goal achievement.) Then the first step to understanding the logical phylogeny of selves requires explaining the logical phylogeny of games amongst socialized but unenculturated biological individuals. Modeling the dynamics by which natural selection can generate and solve non-parametric problems for biological individuals *in general* was the basic founding achievement of evolutionary game theory (Maynard Smith, 1982), so we can treat this step as taken care of.

Now, non-parametric problems are exponentially more complex than parametric ones. This point has been made often, but Sterelny (2004) again offers a nice conceptual extension of it in arguing that non-parametric selection factors make environments 'translucent' to organisms, and in so doing establish selection pressure for representations of some of their features that are both 'robust' and (relatively) 'decoupled'. An organism deploys robust tracking of a feature when its cognitive architecture allows it to represent the feature independently of a specific perceptual stimulus or cue. The architectural conditions for robust tracking have been discussed in the literature for some years. Gould and Gould (1988) offered behavioral evidence of robust tracking in bees. Lloyd (1989) sketched the generic model of a control system he argued to be the minimal requirement for a 'simple mind' because it allows for at least a minimal degree of robust tracking, and there has been empirical discovery of such architectures in cockroaches (Ritzman, 1984), toads (Ewert, 1987) and other animals. Now, robust tracking is required for the implementation of many strategies (which is presumably why it evolved). However, Sterelny argues that humans exhibit to uniquely high degree the use of a representational genus that goes a level beyond robustness in achievement of abstraction. Many representations in humans are *decoupled* from specific action responses. Some of these decoupled representations are standing models of how the world is – beliefs – while others are comparative rankings of ways the world might come to be – preferences. Sterelny defends several interlocked theses concerning decoupled representations: (i) though they occur to some degree in other – invariably social – species, they dominate the ecological life of *H. sapiens* but no other animal; (ii) they are necessary for the *cumulative* transmission of culture; and (iii) though the neural platform that made them possible might have resulted from a rapidly but parametrically changing physical environment during the last series of ice ages, their explosive evolution could only have been driven by the pressures of non-parametric cooperation and competition.

I will here take all three of these theses on board. They are not special to Sterelny, though I know of no one else who has worked out their interrelationships so clearly. Sterelny's key accomplishment with respect to the issues in the present paper is his argument, based on surveyed empirical evidence, that the human capacity for massively decoupled representational scope does not rest on the evolution of special neural mechanisms – though the capacity for robust representation does, and robust representation is necessary for decoupled representation – but on historical

[14] That is, representations that are not tied to any specific class of actions.

dynamics of 'downstream niche construction'. The idea here is that human activity has progressively reconfigured the environment so that (I) there have been steadily increasing returns over time to investment in decoupled representation and (II) the environment is an increasingly efficient storehouse of socially deposited information that cues decoupled representations, which makes them efficient enough for reliable use and which also – crucially – enables developmental processes in children to come to track the eccentric perceptual saliences that decoupling requires, and that other animals largely miss.

Sterelny's various distinctions help us to say a number of things about the nature and role of selves in evolutionary and strategic dynamics. First, we can now be clearer about the relationship between a biological *H. sapiens* individual and a person. The former is a robust representer who instantiates a battery of strategies in various evolutionary games whose players are coalitions of genes. However, the restriction on agent-specific control, at least in its weak form, applies to her, as it does to most animals. Her basic tool in game-playing is her *brain*. She is nevertheless a social animal – manipulation of her parents' responses being the main cognitive behavior for which she must be neurally equipped (Spurrett & Cowley, 2004) – and so the games she must play are relatively informationally demanding (probably similar to those faced by other apes, and by dogs, whales, elephants, pigs, corvids, and parrots). They require a large brain, and this in turn means that potential bottleneck problems in control must be solved at the level of neural design. Neuroeconomics is beginning to tell us how preference control works in the biological *H. sapiens* individual.

For the most part, the current state of play in Santa Fe-style modeling of human economic behavior stops here. These models leave the neuroeconomic details inside black boxes, but that's all to the good. Since the wiring properties on which neuroeconomic facts supervene will change more slowly over time than the selection spaces in which the biological–evolutionary games unfold, as neuroeconomists open the black box this will usefully *constrain* EGT models (by telling us which strategies can and can't be computed) but won't dynamically interact with them *on the same time scale*. Thus the relationship between conventional EGT and neuroeconomics exemplifies the kind of strategy that has worked so well so often in governing the relationships amongst complementary disciplines in the history of science: a 'higher-level' discipline – in this case, evolutionary game theory – isolates and moves around phenomena for a 'lower-level' discipline – in this case, neuroeconomics – to mop up. Then the progressive mopping up feeds back into the higher-level theorizing as *non-dynamic* constraints on possible models. This is great method when you can get it, and people are right to be enthused here.

However, the next thing we learn from reflecting on Sterelny's account is that if we want to understand fully human behavior we can't *stop* with this. A core consequence of the games that our biological *H. sapiens* instance plays is that she will be enculturated into becoming a person. That is, she will be attuned to perceive and be motivated by a range of cultural distinctions and projects that are informationally stored in the ecological relationships between her brain and her environment, rather than in her brain alone. It is very natural to say that a new agent is brought into being by this process. The basic truth in anti-individualism lies right here: this new agent is recruited into existence for the sake of the contributions she can then make to social dynamics. (See McGeer, 2001 for some details on the processes by which human infants are simultaneously led to begin narrating selves and recruited into membership in communities.) Now notice: use of neoclassical formalism will *force* us to say what I have just argued is the natural thing to say. The person will have a different utility function (speaking more properly, a different *sequence of* utility function*s*) from the pre-enculturated biological individual. Indeed, the latter's utility function will range over different goods altogether, since her development involves fundamental re-packaging of perceptual saliences. It is, I suggest, a virtue of the neoclassical formalism that it makes this logical move central and non-optional.

I will now sketch the modeling framework that is implied by all this.

5. Game determination

There is an overlooked puzzle that should have struck game theoretic modelers of human behavior quite independently of the hypotheses about human evolution and development that I have just been discussing. This is the problem I have elsewhere (Ross, 2004, 2005, 2006; Ross & Dumouchel, 2004) called that of 'game determination'.[15] I will reintroduce it here, because it provides independent motivation for the modeling suggestions I'm about to state.

Game theorists building models have a big advantage over people in everyday life (including the game theorists while they're getting on in everyday life). When a game theorist builds a model, she must know, or have justifiably assumed, the utility functions of the players. Her game can *correctly* model a given situation only if her assigned

[15] This construct was originally developed as part of an approach to representing emotional signaling in game-theoretic terms. Emotional response is another phenomenon that has sometimes been thought to confute neoclassical models of human economic behavior, though in this case with *no* basis in the history of fundamental economic thought that will survive serious scrutiny; see Ross and Dumouchel (2004). The confusion seems to rest on a common *double* mistake. First, neoclassical modeling theory gets assimilated with 'rational choice theory', which makes strong empirical assumptions about human behavior that neoclassicism does not (Ross, 2005, 2006, chaps. 3 & 4). Second, muddled critics read 'rational' in 'rational choice theory' as if it were the foil of 'emotional'. This is a folk idea that has sometimes been echoed by philosophers, but has never driven economics. Keynes's famous remarks about 'animal spirits' referred to herding behavior, not emotion. In the classical tradition reason is held to be 'the slave of the passions'. Neoclassical theory incorporates no assumptions at all about the relative weights of different cognitive modalities or styles of behavioral motivation.

utility functions truly describe and predict the players' behavioral dispositions. Of course, most actual game-theoretic models are of stylized or hypothetical agents, because they are investigations of what agents who *did* have such-and-such utility functions (in such-and-such institutional settings with such-and-such information) *would* do. It is because so much game theory activity is of this sort that what I call game determination problems don't loom large in the literature.

Game determination names the task confronting agents who encounter each other, recognize strategic significance in their encounter, but don't know enough about each other's utility functions to be able to know which precise games they might play. This describes most people's situation most of the time. Sterelny's identification of 'translucence' in social environments as the pressure that fueled the evolution of people as cultural niche-constructers is closely related to the point. Determining which games are possible is typically a harder inferential task than modeling or solving a game once utility functions are known. The surest way to keep game determination problems tractable is to build institutions that lock in mutual expectations so long as people are strongly incentivized to want to stay within the institutional rules. Thus groups of mutually anonymous stock market investors, or sales clerks and shop customers, or trade negotiators at the WTO, can get on with their games without having to closely study one another's behavioral idiosyncracies in advance. The need for manageable constraints on game spaces *explains* why cultural rules exist (and then the importance of coordination explains why some particular such rules stabilize). In effect, Sterelny's whole hypothesis that evolution invented culture because social coordination otherwise gets impossibly hard for large-brained organisms in rapidly changing environments is recognition of this point in other terms. Even in the absence of this hypothesis, we can recognize that people somehow solve an information problem in working out which games they're playing when and with whom. The social world doesn't present itself as pre-partitioned into games. This is what I meant by saying that the phenomenon of game determination arises for our attention independently of Sterelny's arguments.

However, these arguments make the problem more pressing. I argued in the previous section that an implication of Sterelny's hypothesis, one that neoclassical (revealed-preference) formalism forces us to stare in the face, is that as biological individuals are enculturated their utility functions change. Furthermore – as Sterelny also emphasizes – they change from being highly predictable (infants all being quite alike[16]) to being relatively distinctive. People learn to be coordinators within *particular* cultures. Their social environments not only make them into German liberals or Alabama fundamentalists or Masai cattle herders; they can even make them into Masai cattle herders who care about model trains but not model airplanes and appreciate British film comedies but not Indian ones. People are capable of inhabiting many cultural niches simultaneously. But characters like Woody Allen's Zelig, who was a perfect cultural chameleon, don't really exist. Again, institutions often smooth things along. Our Masai English comedy buff can figure out how to behave at the Monty Python Fan Club meeting because the group will provide him with all sorts of cues in its newsletters and ceremonial protocols. But he could find himself in deep waters when, out there on the savannah one day, a nature photographer from Texas drives into one of his cows and they have a Situation. The two might face a case of 'radical' game determination.[17]

Our players are not biological individuals; both are enculturated people. Let us therefore denote their game *type* by g'. $g'_{x(i,j)}$ will name a game played by two strangers to each other who are already distinctive human selves. Its structure is of course constrained by their pre-engagement utility functions. But, by hypothesis, they don't know each other's. Crucially, in trying to mutually determine them they are, being people, *bound to act strategically*. In particular, they will strategically signal. So the process of game determination will itself be a game. Also being people, the game playing in which they engage to try to find out which games they might play will amount to further enculturation. The Masai might have no goal of adding a bit of Texan to his cultural repertoire as the negotiation goes on, but because he's a person this might just happen to him nonetheless; and symmetrically for the photographer. Suppose, for example, that they relieve the tension by repairing to the Masai's boma to watch a few old Monty Python episodes on DVD, something of which the American was previously ignorant, but learns to decode and enjoy partly by noticing when the Masai laughs most appreciatively.[18]

Let us put all this first in terms of the narrative theory of the self, then translate that directly into game-theoretic terms. Many engagements among people, where neither detailed mutual personal knowledge nor strict institutional constraints stabilize the dynamics, involve incremental refinements of the selves of the people in question. We might, for useful analytical purposes, build the game $g'_{x(i,j)}$ that *would* describe their play if, like elephants or chimps, they were social but not fully cultural creatures. However, the

[16] Individuals, like almost all animals with brains, vary in *personality*. But personalities, as studied by ethologists, vary along far fewer dimensions than selves do. See Ross (in press) for details.

[17] My phrase here deliberately follows Quine's (1960) 'radical translation'. Quine sought to illuminate everyday problems of meaning interpretation by focusing on a case where two people share no lexical conventions. I similarly am interested in everyday problems of strategic coordination, and find it useful to do so by imagining a case where there is minimal mutual knowledge of shared culturally sculpted utility functions.

[18] Notice how endlessly subtle we can make all this: perhaps the Texan becomes someone who enjoys Monty Python *in something like the way a Masai cattle herder does*. Yes, there can be facts of the matter of this sort. A very experienced anthropologist who studies Masai culture just might be able to spot its influence if she watches the Texan watch Monty Python.

model $g'_{x(i,j)}$ only gives us a baseline from which to *start* modeling the empirical situation, because the people will never actually play $g'_{x(i,j)}$. Instead, they will play another game $g'_{y(i,j)}$ – marked here with the same 'level' indicator ' because it is played by the same players *i*,*j* who 'set out' to play $g'_{x(i,j)}$ – which is distinct from $g'_{x(i,j)}$ because it is played for payoffs over a different set of outcomes. In particular, it is played to determine which game $g_{z(a,b)}$ will be played over the original outcomes of $g'_{x(i,j)}$ (e.g., who shall pay the costs of the dead cow and the wrecked landrover) by the new agents *a*,*b* that are sculpted into being by the play of $g'_{y(i,j)}$.

To bring some analytical order to these complexities, let us define the concept of a 'situation' S that remains invariant through game determination processes. The budget constraints that would have faced the players of $g'_{x(i,j)}$ are inherited by the players of $g_{z(a,b)}$ (i.e., the relative costs of cows and landrovers don't change). Note that this is a stipulation, not something that is empirically guaranteed. 'Deep' re-enculturation *could* change relative costs; imagine the Masai being so charmed by the Texan *lebenswelt* that he 'goes native' and would rather share the repaired landrover than replace his cow. But in designing methodology we get to make practical decisions. We can just stipulate that *if* an invariant situation fails to describe $g'_{x(i,j)}$, $g'_{y(i,j)}$ and $g_{z(a,b)}$ then it's pointless to go on trying to characterize the history using a single dynamic model. Modelers make these sorts of practical decisions, at least implicitly, all the time. Why model the Uruguay Round GATT negotiations as one game and the Doha Round WTO negotiations as another game, instead modeling them as two rounds of one game (with new players joining for round two)? The answer is that too much changes between the rounds for the second option to be sensible. How much change is too much? There surely can be no general a priori rule to govern this judgment call. Suppose, however, that we think that we can get good predictive leverage over $g_{z(a,b)}$ by studying $g'_{x(i,j)}$. (If this were not *often* true then Sterelny's hypothesis would amount to an empirically empty – that is, untestable – metaphysical speculation about the ontogenesis of particular people.) In that case, $g'_{x(i,j)}$, $g'_{y(i,j)}$ and $g_{z(a,b)}$ must all be models of one S_x.

The requirement that we remain within the mathematical rules of game theory imposes tight constraints on our options for interpreting the relationship amongst these models of S_x. *i* and *j* are, in the formalism, strictly different agents from *a* and *b*, but the whole approach here would make no sense if *i* and *j* didn't differentially care about *a* and *b*, or if *i* and *j*, in playing $g'_{y(i,j)}$, were myopic with respect to the predicted equilibria of $g_{z(a,b)}$. The utility functions of *i* and *j*, with respect to the goods up for grabs in $g'_{y(i,j)}$ describe their preferences over which of a range of *g*-level games get played. A *g*-level game is a game amongst players of 'fully determined' games, that is, players who have full information about one another's preferences over the possible outcomes of $g'_{x(i,j)}$. Game theory now gives us two possible options for constraining the solution sets on the games that model S_x.

The first option is what I will call the *minimal constraints approach.* Here, we stipulate that (i) the outcomes over which the utility functions that define $g_{z(a,b)}$ are constructed must include the payoffs available in $g'_{x(i,j)}$; and (ii) the initial state of $g_{z(a,b)}$ is one of the equilibria of $g'_{y(i,j)}$. This approach has the advantage of allowing the modeler flexibility over the degrees to which *i* and *j* predict and are motivated by the utility functions of *a* and *b*. Both levels of strategic myopia and the slopes of discount curves in $g'_{y(i,j)}$ are left as free parameters to be determined empirically (by anthropological study of different sorts of mutual enculturation processes). The price of this flexibility is that the modeling framework isn't doing much work in constraining our accounts, by comparison with seat-of-the-pants situational judgment. This is of course just the standard trade-off one faces in building formal frameworks for representing classes of phenomena.

The second option, the *maximal constraints approach*, would incorporate stronger assumptions into the modeling technology. Here, we would again stipulate the relationship between $g_{z(a,b)}$ and $g'_{x(i,j)}$ just as in clause (i) of the minimal constraints approach. However, clause (ii) will now say that the solution of $g_{z(a,b)}$ must be the subgame-perfect equilibrium of the two-stage game $g'_{y(i,j)} \cup g_{z(a,b)}$ where the terminal nodes state payoffs for each of *i*, *j*, *a* and *b*. Here, *i* and *j* have zero myopia with respect to the welfare of *a* and *b*. (Discount functions remain free parameters.) On this approach, the mutual enculturation described by solving $g'_{y(i,j)}$ is effectively treated as entirely an *informational* transformation.[19]

Note that despite the requirement of subgame perfection, not all coordination signals used in $g'_{y(i,j)}$ need be costly commitment devices. Recent work by Skyrms (2002) shows in detail how use of costless signals can be relevant to reaching equilibria in non-cooperative dynamic games, even if such signals are strategically irrelevant *at* equilibrium. Thus signals that *a* or *b* would not send might be sent by *i* or *j* without this violating the subgame perfection solution and implying that $g'_{y(i,j)}$ is cooperative.

How many real human social exchanges are usefully representable by the maximal constraints approach (by itself) is a strictly empirical question. It will certainly not be fruitful for processes unfolding across generations, or even across major shifts in particular people's maturation cycles or life situations. However, it might provide a powerful source of predictions in application to short-run sequences of encounters. Furthermore, the maximal and minimal constraints approaches could be recursively combined. That is, sequences of games related by maximal constraints could themselves be related to one another by minimal-constraints models. These recursive structures could then be treated as standard formalizations of 'analytic narrative'

[19] Marium Thalos points out that subgame perfection is a very strong equilibrium refinement to impose in a model. That is just one of the factors that makes it appropriate to call the kind of model in question here one that *maximally* constrains the relationships among the elements of S_x.

explanations (Bates, Greif, Levi, Rosenthal, & Weingast, 1998) of medium-run social processes. In any case, I have specified minimal and maximal grades of determination in order to mark off the endpoints on a continuum; many intermediate cases are possible, and these would perhaps be the tools usually appropriate for modeling actual cases. (For instance, one might model a situation as in the maximal constraints case, but replace subgame perfection with Nash equilibrium, with or without concern about trembling hands, etc.)

6. Conclusion

All one can ask of a formalism is that it give one a precise way of stating, and thus of testing, hypotheses. The modeling alternatives I have sketched here do this. To the extent that models close to the maximal constraints end of the continuum of possibilities, used recursively with minimal constraints ones, give us strong predictive leverage over medium-run processes, then game theory will turn out to give us a strongly improved grip on historical–cultural evolutionary change, holding out hope for powerful and precisely formulated generalizations about the patterns of such change. On the other hand, it could turn out that although maximal constraints approaches work well for short episodes within highly stable institutional settings, attempts to chain these together into sequences of minimal constraints models fare little better than traditional historical narratives guided by approximate speculative intuitions about counterfactuals (see Tetlock & Belkin, 1996). (In the modeling methodology described above, this would emerge as inability to reliably induct values of the free parameters in the minimal constraints models from analyses of other models; each minimal constraints model would be built as a largely customized exercise.) In that case, we would be left with the current status quo: standard evolutionary games would help to predict the relative sizes of basins of attraction in long-run games, with the influences of distributions of types of selves left in black boxes; classical games would help to describe short-run interactions among people; and medium-run episodes in which distributions of types of selves matter but these distributions interact dynamically with cultural interaction would resist systematic characterization.

We can capture some of what is at stake here by contrasting the modeling framework I have described with that suggested by Hollis (1998). Hollis argues that people in social interactions often – at least where institutional rules don't explicitly discourage this, as they do in some capitalist markets – strategize by reference to 'team' utility functions that systematically differ from the individual ones they would otherwise manifest. Sugden (2000) argues that this proposal captures some manifest facts about social processes, and Bruni and Sugden (2000) argue that it recovers a classical insight that neoclassicism lost. As a finished account of social dynamics, however, it must leave game theory useless concerning one part of the analysis. In particular, the relationship between team games and short-run games is difficult to itself capture in game-theoretic formalism. Do team games impose commitments on players of short-run games among individuals? If so, what makes these commitments binding? In effect, Hollis's proposal makes games amongst individual people into stages of larger *cooperative* games. Two unwelcome consequences follow from this. First, it amounts to supposing that cultural pressure for cooperativeness has managed to completely swamp (at least in normal cases) natural-selection pressures that encourage *biological* individuals to compete with one another. This isn't impossible, but it is a very strong hypothesis. It immediately implies the second problematic feature of the framework, which is that, once again, distinctive selves become causal epiphenomena; individual people are just robotic products of team dynamics.

By contrast, on the framework I have suggested here social dynamics *are* logically and ontogenetically prior to individual selves, because selves are sculpted into being by social processes. However, the outcomes of *g* level games – the actual bargaining episodes that determine the particular distributions of strategically created and contested assets – are sensitive to the properties of individual narratively generated selves. Furthermore, properties of biological individuals are inputs to the social processes. The framework imposes no a priori view on the relative causal strengths of Darwinian competition amongst biological individuals and the stabilization of cooperative dispositions under the evolution of institutionalized norms. Finally, no part of the process fails to be describable in the standard formalism of game theory, as interpreted by reference to Samuelsonian neoclassical preference theory. The framework thus satisfies all the desiderata developed in the earlier sections of the paper. Whether it will help us to predict and explain empirical phenomena that otherwise resist systematic treatment must remain to be seen.

Acknowledgements

I thank Andy Clark, Stephen Cowley, Harold Kincaid, David Spurrett, two anonymous referees for this journal, and audiences at 'Collective Intentionality' (Sienna, 2004) and 'Distributed Cognition and the Will' (Birmingham Alabama, 2005) for their comments on earlier drafts of this paper.

References

Ainslie, G. (2001). *Breakdown of will*. Cambridge: Cambridge University Press.

Bates, R., Greif, A., Levi, M., Rosenthal, J.-L., & Weingast, B. (1998). *Analytic narratives*. Princeton: Princeton University Press.

Binmore, K. (1998). *Game theory and the social contract. Just playing* (Vol. 2). Cambridge, MA: MIT Press.

Binmore, K. (2005). *Natural justice*. Oxford: Oxford University Press.

Bouchard, T., & Loehlin, J. (2001). Genes, evolution and personality. *Behavior Genetics, 31*, 243–273.

Bowles, S. (2004). *Microeconomics: Behavior, institutions and evolution*. Princeton: Princeton University Press.

Bruner, J. (1992). *Acts of meaning*. Cambridge, MA: Harvard University Press.
Bruner, J. (2002). *Making stories: Law, literature, life*. New York: Farrar, Strauss and Giroux.
Bruni, L., & Sugden, R. (2000). Moral canals: trust and social capital in the work of Hume, Smith and Genovesi. *Economics and Philosophy, 16*, 21–45.
Buss, L. (1987). *The evolution of individuality*. Princeton: Princeton University Press.
Camerer, C., Loewenstein, G., & Rabin, M. (Eds.). (2003). *Advances in behavioral economics*. Princeton: Princeton University Press.
Clark, A. (1990). Connectionism, competence and explanation. In M. Boden (Ed.), *The philosophy of artificial intelligence* (pp. 281–308). Oxford: Oxford University Press.
Clark, A. (1997). *Being there*. Cambridge, MA: MIT Press.
Clark, A. (2002). That special something. In A. Brook & D. Ross (Eds.), *Daniel Dennett* (pp. 187–205). New York: Cambridge University Press.
Clark, A. (2004). *Natural born cyborgs*. Oxford: Oxford University Press.
Davis, J. (2003). *The theory of the individual in economics*. London: Routledge.
Dennett, D. (1987). *The intentional stance*. Cambridge, MA: MIT Press.
Dennett, D. (1991). *Consciousness explained*. Boston: Little Brown.
Dennett, D. (2003). *Freedom evolves*. New York: Viking.
Elster, J. (1985). *Making sense of Marx*. Cambridge: Cambridge University Press.
Elster, J. (1999). *Alchemies of the mind*. Cambridge: Cambridge University Press.
Ewert, J.-P. (1987). Neuroethology of releasing mechanisms: prey-catching behavior in toads. *Behavioral and Brain Sciences, 10*, 337–368.
Flanagan, O. (2002). *The problem of the soul*. New York: Basic Books.
Gintis, H. (2000). *Game theory evolving*. Princeton: Princeton University Press.
Gintis, H. (2004). Towards the unity of the human behavioral sciences. *Politics, Philosophy and Economics, 3*, 37–57.
Glimcher, P. (2003). *Decisions, uncertainty and the brain*. Cambridge, MA: MIT Press.
Gould, J., & Gould, C. (1988). *The honey bee*. San Fransisco: Freeman.
Hodgson, G. (2001). *How economics forgot history*. London: Routledge.
Hollis, M. (1998). *Trust within reason*. Cambridge: Cambridge University Press.
Hull, D. (1976). Are species really individuals? *Systematic Zoology, 25*, 174–191.
Kennedy, J., & Eberhart, R. (2001). *Swarm intelligence*. San Fransisco: Morgan Kauffman.
Keller, E. F. (1999). *The century of the gene*. Cambridge, MA: Harvard University Press.
Lawson, T. (1997). *Economics and reality*. London: Routledge.
Lloyd, D. (1989). *Simple minds*. Cambridge, MA: MIT Press.
Maynard Smith, J. (1982). *Evolution and the theory of games*. Cambridge: Cambridge University Press.
McGeer, V. (2001). Psycho-practice, psycho-theory and the contrastive case of autism. *Journal of Consciousness Studies, 8*, 109–132.
Montague, P. R., & Berns, G. (2002). Neural economics and the biological substrates of valuation. *Neuron, 36*, 265–284.
Ritzman, R. (1984). The cockroach escape response. In R. Eaton (Ed.), *Neural mechanisms of startle behavior* (pp. 93–131). New York: Plenum.
Roemer, J. (1981). *Analytical foundations of marxian economic theory*. Cambridge: Cambridge University Press.
Rosenberg, A. (1992). *Economics: Mathematical politics or science of diminishing returns?* Chicago: University of Chicago Press.
Ross, D. (2002). Why people are atypical agents. *Philosophical Papers, 31*, 87–116.
Ross, D. (2004). Meta-linguistic signaling for coordination amongst social agents. *Language Sciences, 26*, 621–642.
Ross, D. (2005). *Economic theory and cognitive science: Microexplanation*. Cambridge, MA: MIT Press.
Ross, D. (2006). Evolutionary game theory and the normative theory of institutional design: Binmore and behavioral economics. *Politics, Philosophy and Economics, 5*, 51–79.
Ross, D. (in press). *H. sapiens* as ecologically special: what does language contribute? *Language Sciences*.
Ross, D., & Dumouchel, P. (2004). Emotions as strategic signals. *Rationality and Society, 16*, 251–286.
Rothschild, E. (2001). *Economic sentiments: Adam Smith, Condorcet, and the enlightenment*. Cambridge, MA: Harvard University Press.
Samuelson, P. (1947). *Foundations of economic analysis*. Cambridge, MA: Harvard University Press (Enlarged edition, 1983).
Satz, D., & Ferejohn, J. (1994). Rational choice and social theory. *Journal of Philosophy, 91*, 71–87.
Schelling, T. (1980). The intimate contest for self-command. *Public Interest, 60*, 94–118.
Skyrms, B. (2002). Signals, evolution and the explanatory power of transient information. *Philosophy of Science, 69*, 407–428.
Spurrett, D., & Cowley, S. (2004). How to do things without words: infants, utterance-activity and distributed cognition. *Language Sciences, 26*, 443–466.
Sterelny, K. (2004). *Thought in a hostile world*. Oxford: Blackwell.
Stigler, G., & Becker, G. (1977). De gustibus non est disputandum. *American Economic Review, 67*, 76–90.
Sugden, R. (2000). Team preferences. *Economics and Philosophy, 16*, 175–204.
Sugden, R. (2001). The evolutionary turn in game theory. *Journal of Economic Methodology, 8*, 113–130.
Tetlock, P., & Belkin, A. (Eds.). (1996). *Counterfactual thought experiments in world politics*. Princeton: Princeton University Press.
Tomasello, M. (2001). *The cultural origins of human cognition*. Cambridge, MA: Harvard University Press.
Weibull, J. (1995). *Evolutionary game theory*. Cambridge, MA: MIT Press.
Young, H. P. (1998). *Individual strategy and social structure*. Princeton: Princeton University Press.