CORA Cork Open Research Archive
Cartlann Taighde Oscailte Chorcaí

Title	Economics, social neuroscience, and mindshaping
Authors	Ross, Don;Stirling, Wynn
Publication date	2020-09-24
Original Citation	Ross, D. and Stirling, W. (2020) 'Economics, social neuroscience, and mindshaping', in Harbecke, J. and Herrmann-Pillath, C. (eds) Social Neuroeconomics: Mechanistic Integration of the Neurosciences and the Social Sciences. London: Routledge, pp. 174-202.
Type of publication	Book chapter
Link to publisher's version	https://doi.org/10.4324/9780429296918, https://www.routledge.com/Social-Neuroeconomics-Mechanistic-Integration-of-the-Neurosciences-and/Harbecke-Herrmann-Pillath/p/book/9780429296918
Rights	© 2021, the Authors. This is an Accepted Manuscript of a book chapter published by Routledge/CRC Press in Social Neuroeconomics: Mechanistic Integration of the Neurosciences and the Social Sciences on 23 September 2020, available online: http://www.routledge.com/9780429296918
Download date	2026-10-06 18:43:25
Item downloaded from	https://hdl.handle.net/10468/11077

UCC
University College Cork, Ireland
Coláiste na hOllscoile Corcaigh

Economics, Social Neuroscience, and Mindshaping

Forthcoming in Jens Harbecke and Carsten Herrmann-Pillath, eds., *Social Neuroeconomics: Mechanistic Integration of the Neurosciences and the Social Sciences*. Routledge, 2020. Don Ross

don.ross931@gmail.com
School of Sociology, Philosophy, Criminology, Government, and Politics, University College Cork
School of Economics, University of Cape Town
Center for Economic Analysis of Risk, Georgia State University

Wynn Stirling
wynn_stirling@byu.edu
Department of Electrical and Computer Engineering, Brigham Young University

Abstract

We consider the potential contribution of economics to an interdisciplinary research partnership between sociology and neuroscience ('social neuroscience' or 'social neuroeconomics'). We correct a misunderstanding in previous literature over the understanding of humans as 'social animals', which has in turn led to misidentification of the potential relevance of game theory and the economics of networks to the social neuroscience project. Specifically, it has been suggested that these can be used to model mindreading. We argue that mindreading is at best a derivative and special basis for social coordination, whereas the general and pervasive phenomenon on which it depends is *mindshaping*. We then outline the foundations of Conditional Game Theory as a mathematical model of mindshaping, which extends game theory without displacing its classic solution concepts, and which exploits economists' experience in modeling networks.

JEL codes: A11, A12, C73, D85, D87, D91, Z13

1 Introduction

The past few decades have witnessed increasing entanglement of economic and psychological research, now appearing routinely in the leading economics journals. Some of this, under the banner of neuroeconomics, has encompassed work on the computation of value in the brain, and on invocation of specific neural mechanisms as sources of bounds on idealised rationality (Ross 2008). Economists have also been increasingly open to borrowing insights from sociology (Akerlof and Kranton 2010; Frijters and Foster 2013), and Ross (2014) argues that the scope for integration along that disciplinary frontier is ultimately deeper than on the borderlands with psychology. In this context, recent efforts to build a research programme of sociological neuroscience (Schutt, Seidman, and Keshavan 2015)[1] open a new and potentially intriguing route for cosmopolitan economic methodology.

One of the main bases for skepticism about neuroeconomics has been its implicit reductionism. Neuroeconomists, at least in their rhetoric, tend to equate multiply realizable equivalence classes of choices with particular mechanisms that might sometimes constitute realizers (Fumagalli 2014, 2016). Ross (2014) mounts a wider criticism, that many neuroeconomists implicitly try to transform economics into the psychology of individual valuation, choice and behaviour, thereby underweighting the discipline's history and

[1] We refer to this programme as 'new' because previous social neuroscience, as gathered in Cacioppo, Visser, and Pickett (2006) is not explicitly integrated with sociological research.

applications that are more consistent with the aims and objectives of a social science. Economics, Ross argues, is primarily concerned with strategic interactions of agents in institutional and cultural environments, not with individual decisions abstracted from 'confounding' social influences.

A new way of responding to this criticism might involve incorporating neuroscientific results and models within the context of a tridisciplinary mélange that includes sociology.[2] Alós-Ferrer (2018) chides Schutt et al (2015) for failing to invite economists to the neurosociology party, and goes on to identify some forms of expertise that economists can potentially contribute to the neuroscience-sociology relationship. Specifically, he suggests that whereas neuroeconomics has amounted to a study of implementation processes for individual decision theory, a modeling approach already shared with psychologists, social neuroscience (or, alternatively, 'social neuroeconomics') invites economists to deploy their game theoretic toolbox. Alós-Ferrer additionally observes that economics is the social science that has generated the most sophisticated applications of mathematical network theory (Goyal 2007; Ioannides 2013), and this is another natural technology for social neuroscientific modelers.

We concur with both of these suggestions. However, we argue that Alós-Ferrer is right for wrong reasons, and that he consequently recommends *mis*application of both game theory and network theory to social neuroscience. His mistake rests on misinterpreting anthropological evidence about the evolution of human sociality, and on ignoring leading recent themes from the philosophy of cognitive science. The significance of this criticism lies in lessons it offers for a methodology of economically inflected social neuroscience.

2 Individualism and human sociality

Early in their volume Schutt et al (2015) include an essay by Jonathan Turner (2015) on the evolutionary anthropology of humans. Alós-Ferrer (2018) summarises it at length in his review, endorsing but also polemically extending Turner's interpretation of anthropological and neuroscientific evidence.[3] In most respects Turner's narrative of the natural history of *H. sapiens* is relatively uncontroversial. At its centre is the fact that great ape group formation follows the fission-fusion pattern, in contrast to the large, stable matrilineal clans of monkeys. This is attributed, again plausibly, to monkeys' access to more reliably and densely distributed food sources in the middle regions of forest canopies about 23 million years ago, while apes foraged in less ecologically friendly parts of the forest ecosystem.[4] Then, about 14 million years ago apes, who had become smarter than monkeys under pressure from their more precarious environment, were forced onto the open savannah by climate change. This required them to form more cohesive groups for hunting game and for defense against cats and hyenas, but they needed to do so using brains that had adapted to the fission-fusion social lifestyle. Development of language was ultimately part of the neuroadaptive response in the case of hominins, but more fundamental, according to Turner, was growth of brain areas that enhanced emotional response, including emotions associated with social bonding. Along with these prosocial emotions came amplified expressions of fear, and the various nuances of social aversion on which the attention of human artists recurrently lingers. Love of specific others supports bonds that can be quite powerful, but also tend to be unstable. Artists have been particularly fascinated by the dynamics that shift human love to hatred (and, less frequently, hatred to love). Based on both behavioural and neural evidence, if baboons and capuchins could read human novels they would not be able to empathise with such narratives.

[2] Alós-Ferrer (2018) notes that this will require reduced reliance on fMRI studies relative to the standard practice in neuroeconomics. Preoccupation with fMRI has saddled neuroeconomics with serious statistical inference problems that econometricians have noticed more intently than other parts of the critical audience (Harrison 2008). Alós-Ferrer (2018) helpfully discusses and recommends alternative methods for social neuroscience.

[3] We refer to Alós-Ferrer's extensions as 'polemical' both because they are more controversial than Turner's own claims, and because they are not supported in his text by citations of additional evidence.

[4] Turner conjectures that this resulted from monkeys' advantage in being able to digest unripe fruit.

Being, as Turner puts it, "bioprogammed" for cohesion, they do not experience the kinds of social drama over which humans obsess.

When Alós-Ferrer (2018) compresses this post-arboreal history and foregrounds the aspects he thinks will be most significant for his fellow economists, he describes it as a flawed transition from "individualism" to socialization. Going beyond Turner, his rhetoric emphasizes that it is the opposite story from the one that economists (e.g. Ofek 2001, Ross 2013) typically tell when they consider deep history. These latter accounts describe mechanisms by which naturally tribal normative conformists came to culturally construct and institutionally protect forms of individualism that encourage specialisation of labour.[5] For non-amateur tellings of this tale, i.e. by historians rather than economists, see Morris (1972) and Taylor (1989). But Alós-Ferrer overtly means to upset apple-carts here:

> ...the oft-spouted claim that humans are social animals is, simply put, at odds with all available evidence, and seems to arise from "wishful thinking" related to cultural (Western) values. This is a fundamental realization for social neuroscience *and* economics. Of course, the human brain has adapted to facilitate social interactions. But this is a forced and problematic adaptation, which goes against a previous adaptation that took away social group-formation tendencies. It does not take much to expect trouble arising from the "social brain", since, in evolutionary terms, it amounts to a recent patch on top of an individualistic brain [2018, p. 248].

This apparent direct conflict between two narratives - one leading from natural individualism to 'flawed' sociality, the other tracing development from conformism to market-adapted individualism - is merely apparent, because they foreground different timescales in human history. Turner and Alós-Ferrer compare Miocene apes with hominins. Economists such as Ofek, and Ross, and historians such as Morris and Taylor, compare earlier and later *people*. In consequence, the meaning of 'individualism' shifts between the two broad narratives that Alós-Ferrer frames as rivals. Turner refers to "strong ties" among the monkeys he contrasts with apes, rather than to any straightforward antonym to "individualism" as the latter has been understood in political economy. What is meant by 'strong ties' are associations that are relatively less likely to break under stress. To the extent that such ties permeate a whole troop of monkeys, the troop will be less fissile. In Turner's story of comparative adaptations, this extends to tactical behavioral patterns: faced with a threatening lion, a baboon troop keeps its defensive formation and the big males who are crucial to defense do not separately panic. Turner speculates that less reliable cohesion in such encounters explains why most ape species went extinct and why none but humans still live in open savannah. Then the anthropologically informed sociologist, wondering how humans managed to re-forge reliable enough coordinated response to get through their evolutionary challenge when they descended onto the plains, seeks mechanisms that could support such unflinching whole-group stability.

Economists such as Ofek (2001) and Ross (2013) have a quite different explanandum in view: conditions and mechanisms that support division of social roles and of labour. More specifically, they focus on sustained incentives and dispositions that motivate individuals to seek competitive advantages for themselves through behavioural and technological innovations while retaining enough mutual understanding with others to make social learning possible.

Alós-Ferrer's additions to Turner's account suggest equivocation around these different explananda. Whereas Turner stresses the complex relationship between stormy emotions and socially sanctioned self-control in the maintenance of social order, Alós-Ferrer calls this part of his account "incomplete", and emphasises that "along the way, something amazing happened. The great apes started developing the capacity to *mentalize*, that is, the capacity to understand other individuals' minds" (2018, p. 249). He goes

[5] Martens (2004) is particularly insightful on the psychological and strategic relationships between individuality and specialisation of labour.

on to suggest that mindreading, the conjecturing of frequently successful theories about unobserved 'inner' preferences and beliefs of others, is the core capacity that many commentators take to be formalised in applications of game theory.[6] Alós-Ferrer effectively views the capacity to mindread as an evolutionary kludge that replaces lost 'bioprogramming' for social cohesion. One might gloss Alós-Ferrer's extension of Turner by appeal to game theory as follows: whereas mindreading allows humans to play sophisticated *non-cooperative* games, 'bioprogramming' causes monkeys to face corresponding social challenges and opportunities as *cooperative* games.

It is noteworthy that in their accounts of human historical socio-cognitive milestones neither Turner nor Alós-Ferrer mention storage of representations in the external, socially observable, environment. This is a potential basis for an alternative narrative about the special nature of human sociality, one that is widespread in the literature on the deep origins of human culture (and that is also discussed extensively in Herrmann-Pillath's chapter in the present volume). Under this framing, humans *are* uniquely "social" animals, notwithstanding their tendencies to develop emotional animosities and to free ride in responding to threats to the group.[7] A human, to be biologically, socially, and economically successful, is obliged to devote continuous and careful attention to representations of reports of facts, hypothetical conjectures, and normative injunctions generated by other humans. As various writers (see especially Sterelny 2003) have stressed, the history of *H. sapiens* is in large part the story of cognitive and behavioural adaptation to an environment in which the dominant ecological variables are cultural. As Sterelny (2012) emphasises, this massively expanded the extent and importance of deliberate pedagogy in humans, along with capacities for strategically promulgating and detecting falsehoods, fantasies, and normative propaganda. That neither Turner nor any other contributor to Schutt et al (2015) mentions shared external representation is a surprising oversight in a volume surveying relationships between sociology and neuroscience.

As noted, Alós-Ferrer (2018) draws attention to what he regards as a different gap, that the book includes no discussion of the human capacity for mentalising in general, and specifically for mindreading. As an aid to his review of topics covered in Schutt (2015), Alós-Ferrer (2018, p. 239) provides a diagram of hypothesised networks in the brain for processing information about, and potential actions over, various social domains. He includes among these a "Theory of Mind" network located in the dorsal-medial prefrontal cortex and ventral-medial prefrontal cortex, and urges that economists can contribute to the general social neuroscience project by modeling the operations of this network using game theory.

We concur with Alós-Ferrer that the cognitive, as opposed to emotional, dimension of human sociality is severely under-represented in Shutte et al (2015). However, we argue that Alós-Ferrer's proposal as to how economists might best step in to pick up this slack is essentially wrong-headed, and that this reflects his misinterpretation of the anthropological record as implying that humans are fundamentally individualistic, in a sense that classical applications of game theory generally reflect. We go on to argue that game theory *can* be adopted to apply to humans as the kind of social animal that, pacé Alós-Ferrer, they really are.

3 Mindreading

Turner's emphasis on emotions as social regulators follows a strong trend in neuroscience that found its most influential rallying call in Antonio Damasio's (1995) classic book *Descartes' Error*. Though Damasio wrote for a general educated audience, his book influenced professional researchers because, in identifying

[6] This is not the universal interpretation among economists of the role of game theory as a technology for modeling behaviour. It is, however, the view of the large number of researchers who think that game theoretic accounts, to be empirically valid, must be at least approximately correct descriptions of actual processes of strategic reasoning by players to which they are applied. Other theorists, including us, prefer to view games as mathematical models for predicting outcomes of multiply realizable types of strategic scenarios, from which the economist deliberately abstracts (Ross 2014). On this view game-theoretic applications should not generally be interpreted as modeling reasoning processes.

[7] Again, it is Alós-Ferrer, not Turner, who associates these problems with "individualism".

what he saw as a gap between the neuroscientific study of the emotional brain and an unduly rationalistic cognitive science, Damasio sought to transcend limitations of both specialist literatures.[8] It is noteworthy that Damasio chose Descartes as his foil, because Descartes's rationalism was intimately bound up with his atomism. In his *Meditations* Descartes famously tried to provide arguments against epistemic solipsism, the concern that no one can know on the basis of observation that any consciousness exists other than her own. In the opinion of most philosophers Descartes's arguments were not successful, leaving their discipline with a 'problem of other minds' on which they have been chewing ever since.[9] Because Descartes viewed emotions as corporeal and cognition as the essence of the mental, he certainly could not have proposed emotional resonance as an answer to the question of how people know about one another; this is part of the excessive rationalism that Damasio takes as his rhetorical opponent.

The large philosophical literature on mindreading is heir to the Cartesian tradition, though modern philosophers are typically willing to allow that emotions are important features of minds, and even that some emotions (e.g. empathy) may play a role in facilitating mindreading. Philosophers are divided over whether people 'read' minds by literally conjecturing and testing folk psychological theories about them (Carruthers 1996), or whether they base their projections on simulations they run using their own decision-making and reasoning equipment. So-called 'simulation theorists' (e.g. Goldman 2006) were inspired by the discovery of so-called 'mirror neurons' in pre-frontal cortex (Rizzolatti, Fogassi, and Gallese 2001), which they generally interpreted as (at least) contributors to the simulations they hypothesise. (Since mirror neurons were initially observed in monkeys, this history comports awkwardly with the suggestion that monkeys 'bioprogrammed' for social coordination don't need to mindread.) In the spirit of Hegel the majority of philosophers seem lately to be coming around to a hybrid of the two kinds of account (Nichols and Stich 2003).

The most important empirical evidence that is cited in favour of mindreading comes from developmental psychology. Children under the age of 4 years are not able to reliably recognize that others have perspectives different from their own, but (it is claimed) without explicit training they naturally begin ascribing idiosyncratic beliefs and desires to others after that milestone (Wellman 1990). In the context of the Piagetian tradition, this has been interpreted as the onset of a cognitive-inferential capacity that is a natural disposition even if it might need to be triggered by exemplars in the social environment. It is not controversial that children require a minimum level of cognitive, and possibly linguistic, sophistication before they can frame non-egoistic concepts of beliefs and desire, which philosophers call 'propositional attitudes' (PAs) because they involve general operations over an unlimited set of potential descriptions of states of affairs (i.e., one can believe that x or desire that y where x and y are drawn from relatively unrestricted sets of descriptions of the world). Philosophers disagree over whether a child should be said to genuinely understand PAs during the stage at which they only express them egoistically. A tradition of experiments by psychologists has focused on whether children are surprised by violations of observable perspective asymmetry before they can articulate PAs, or indeed can speak at all.

We need not question the developmental evidence to wonder whether what children start doing when they begin attributing reasonable PAs to other people (and to dogs and dolls and imaginary agents) amounts to using a theory (or simulations) to predict what they will do. Clearly this involves *interpretation* of the observed behaviour.

We referred to mindreading theorists as heirs to Descartes because they share the general philosophy of mind that informs classic Western epistemology. This is that PAs are intended to refer to real states of

[8]There is a general tendency among academics interested in interdisciplinary relationships, including philosophers trying to explicate the unity of science, to pay too little attention to serious popular science. This no doubt reflects a widespread feeling that it is a bit grubby to engage with and cite such work. But it is precisely the popular science literature that, in trying to satisfy widespread demand for a general scientific worldview that can partly plug the hole in human life left by the decline of religious metaphysics, also directs researchers' attention to potential connections between their own fields and others.

[9]Non-philosophers can find a clear and compact outline of this (arguably perverse) problem in Churchland (1988), pp. 67-72.

people's minds.[10] According to the tradition, these states are *internal*; this is why they must be inferred, either through use of general-purpose cognition or by means of simulation, rather than observed. Economists should be able to grasp the philosophical picture without needing detailed examples, since it is an elaboration of everyday Western folk psychology. More relevantly, economists are familiar with constructing Bayesian games, in which a Player X_i who is uncertain about another player X_j's preferences over a set of outcomes Θ infers $\mu_j(\Theta)$ statistically from information revealed through game play, which X_i might strategically try to elicit by taking screening and signaling actions (e.g. Misyak et al 2014). Some economists will balk at the idea that people literally process Bayesian probabilities when they are involved in interactions modeled as games under uncertainty, but many will not; and even the former will likely grant that whatever players of Bayesian games are taken to be up to psychologically, it is some form of cognitive inference. We infer from his remarks about the role that game theory might come to play in social neuroscience that Alós-Ferrer is among the economists who are internalists and realists about PAs, and who think that preferences and beliefs included in game-theoretic models are intended to at least roughly correspond to them.

Increasingly many philosophers, however, have broken with the Cartesian heritage, and to a more radical extent than Damasio or the social neuroscientists. Over the past three decades, a majority have been persuaded to reject internalism about PAs. This shift has been a primary motivator for recent expressions of skepticism about the importance of mindreading to social cognition and coordination.

4 Externalism about propositional attitudes

Some early approaches to artificial intelligence (e.g. Newell and Simon 1976) amounted to implementations of the hypothesis that a mind can function in real time by directly performing computations over literal internal encodings of PAs. As an empirical hypothesis about the human mind, this quaintly straightforward idea, according to which folk psychology is a valid construct from direct, veridical inner observation, was systematically elaborated and defended by some theorists (Fodor 1975; Pylyshyn 1984); for a brief period it was the dominant model. However, it provoked skeptical critics from the outset. Two whose (respective) articulated grounds for doubt have been closely vindicated by subsequent science were Dennett (1969) and Dreyfus (1979). There is not space here to rehearse these criticisms or their relationships to less naïve research programmes in AI that succeeded them. The point of mentioning this history is to remind readers that the idea that PAs are found discretely stored in brains, or have direct representational isomorphs in brains, has never been treated in cognitive science as the common-sense default position that folk psychology, and by extension most economists, frequently seem to take it to be. In AI it was a bold, ambitious hypothesis, which was abandoned relatively quickly as being both biologically implausible and technologically impractical.[11]

Logically, if PAs are not 'in people's heads' then either they must be grounded at least partly *outside* of people's heads or they must be regarded as fictitious constructs. Two decades of intense debate among philosophers explored the implications of, and evidence for, both interpretations. This literature is difficult for non-philosophers to profitably visit, because much of it was preoccupied with the metaphysics of symbolic reference and with questions about how the concept of knowledge, with its requirement of objectively accurate representation, might be preserved if commitment to the interpretation-independent reality of internally represented PAs were given up.[12] But by the end of the 1990s consensus was emerging, among philosophers of mind and among those psychologists who didn't prefer to pragmatically skirt the issue, that

[10] Some philosophers, e.g. Churchland (1981) believe that these intended references fail, that is, that ascribed PAs are *mis*characterisations of internal states that should ultimately be discarded as science reveals more precise and neuroscientically grounded states that will yield more accurate predictions than folk psychology.

[11] These objections were developed in AI under the label of 'the frame problem'; see Pylyshyn (1987).

[12] This preoccupation leads philosophers to distinguish between narrowly 'semantic' externalism and externalism about the empirical basis for successful cognitive management of environmental contingencies. We are concerned only with the latter.

PAs are social constructs that facilitate, or are indeed essential for, coordination of communication and practical interaction among people. 'Social constructs' tends to be read outside sociology as implying fictional status. The now standard view that philosophers call 'externalism' typically avoids this connotation by referring instead to 'virtual reality', with the implication that being virtual is a way of being *real* (see Dennett 1981).

It is somewhat puzzling that, particularly in the current technological environment, externalism still seems to be regarded outside cognitive science as an exotic view. It has scarcely penetrated economics at all, though it is implicitly present in some methodological reflections of Vernon Smith (2008) (see Dekker and Remic 2019), and its potential relevance to economists is explored in detail by Ross (2005, 2014) and Herrmann-Pillath (2013). The proposition, comprehensively surveyed in McClamrock (1995),[13] is that PAs are *ascriptions* to people of dispositions to behave, to make inferences, to communicate, and to signal emotions in ways that simultaneously (a) *summarily consolidate* information to which they are cognitively and conatively responsive, (b) *rationalise* patterns of action and communication so to render them coherent and mutually reinforcing, and (c) *predict* ranges of expected future behaviour including self-descriptive behaviour. A central idea of externalism is that a necessary aspect of social competence is that a person fluently and regularly ascribes PAs to *herself* that renders her intelligible to others *and* to herself *as* a member of a network of interpreters. A person, on this account, does not observe or infer her private representational states and then report her findings about herself to others. Rather, she *uses* the resources for self-characterisation that her society, and its language, make available to her to dynamically and incrementally *construct* a self that is relatively unified and consistent in action over time and across equivalence classes of incentives, available information, and prospects for improved flourishing. Indeed, she is normatively *obliged* to to do this as a condition on participating in joint projects with others who must be able to project (relatively) stable expectations.

PA ascription can perform its fundamental role of stabilising and coordinating expectations among people only to the extent that there are general patterns of agreement on which beliefs and desires best rationalise observed patterns. We will consider the mechanism for this in more detail in the next section. A crucial enabling resource for stabilisation, among people, is semiotic *scaffolding* in the shared, constructed environment. If I know that you are planning to go to an institution that we both recognize, through culturally established signs, as a bank, and I in addition see that you are carrying a brochure about business start-up loans, my set of plausible conjectures about which beliefs and desires might most economically rationalise your behaviour is sharply constrained. The most important source of scaffolding is a shared language itself, which forces convergence on a common general ontology of categories of objects, processes, roles, and labels for motivations.

A skill learned by all socially competent people is to *prepare* for exchanges that involve complex or idiosyncratic motivations and intentions by self-ascribing PAs in a self-narrative and then bringing the immediate visible semiotic environment into alignment with the narrative in question. When a person does this she is not, contrary to the internalist idea, *discovering* relevant linguistically encoded messages lurking in her brain through introspection. There are no sentences in her brain until and unless she literally puts acoustically and linguistically structured memories there by deliberately, silently speaking in English to herself. But this is something people routinely do every day; they learn what they think when they hear what they say, including to themselves.

Externalism thus represents recognition by cognitive scientists that PAs are mappings between processes in brains that are mainly beyond direct observational and conventionally describable access and practically focused, normatively regulated and socially constructed ontological coordination grids encoded in the exter-

[13] We cite a preferred source; there are many other general accounts. Externalism developed incrementally through contributions by numerous theorists, and we are aware of no plausible claim to priority by an author. The most cited externalist treatise is Clark (1997), but he is more concerned to extend and apply the view than to motivate and explain it.

nal (social) environment. It differs from the internalism that the early AI theorists tried to implement – and that got Descartes stuck trying to argue his way out of solipsism – mainly in holding that that's *all* that PAs are; they are not public copies of private, internal beliefs and desires 'written' in 'brain code'. The doctrine is thus deflationary in *a sense*, but it hardly renders PAs of derivative or downgraded importance: a person could not participate in society, and thus could not survive as a person, without having learned how to use them. Beliefs and desires are *social tools*, a point well articulated and defended by Pesonen (this volume).

One might think that a view of PAs according to which they are irreducibly social entities would be appealing to sociologists, including neurosociologists. Again, however, there is little evidence that externalism has penetrated the social sciences outside of some of its newer engineering branches, e.g. the discipline of information systems (Clark 2003). Where social neuroscience is concerned, this may reflect the relative neglect of cognitive interaction in favour of emotional communication and influence, since emotions are more typically expressed directly than reported indirectly using PAs.

We earlier indicated our agreement with Alós-Ferrer (2018) that social neuroscience should pay much more attention to cognitive phenomena than is evident in the Schutt et al (2015) volume. When this correction is made, one would hope that the externalist perspective will be picked up. We have not, however, identified any specific problems threatening the agenda of social neuroscience based on the absence, so far, of this perspective. We have explained externalism here as a stepping stone to a more recent critique that is partly based upon it, of the importance currently attached to mindreading in cognitive science. As described earlier, it is with respect to modeling mindreading that Alós-Ferrer sees a special role for economists, particularly game theorists, in social neuroscience. We now explain why we think this is wrong.

5 Mindshaping

Externalism about PAs is not strictly incompatible with the hypothesis that humans coordinate with one another by mindreading. There might be occasions when a person has reason to try to infer what another has said to herself - for example, in a game of charades. But if we are persuaded by externalism then this is hardly likely to strike us as a good model of the typical case. From the externalist perspective, the mindreading hypothesis seems to get the dependence relation between PA ascriptions to oneself and PA ascriptions to others backwards. People talk to themselves, thus creating 'private' PAs, when they have reason to pre-rehearse a specific conversation they anticipate with another person, as well as when they are uncertain about which propositions their evidence warrants or are unsure what they want. In the normal course of life, however, it is the stream of public PAs, co-created by interacting dyads or larger groups of interactants, that are primary. People's models of themselves in terms of PAs are based on the models that others use to make sense of their actions and utterances. By contrast, emphasis on mindreading seems to presuppose that we start from well-articulated self-models that others then work to figure out.

The earliest explicit challenges to the mindreading hypothesis were based on observations of general human problem-solving strategies, and on the interactive nature of human parenting. Clark (1997) observed that people often respond to an initially intractable problem not by inventing new solution tactics but by manipulating the problem so as to make it more amenable to familiar methods. Why would they not solve social coordination challenges in the same way? This would involve trying to change the behaviour of others rather than concentrating on accurately modeling them for the mere sake of entertaining correct theory. It will be recalled that an important source of the mindreading hypothesis was developmental psychology. The studies on which these analyses were based tended to study children placed in situations where they were called upon to answer questions about novel situations using only their 'inboard' cognitive resources. But young children develop through intensive interaction with parents and other teachers. Mameli (2001) and McGeer (2001) independently drew attention to the fact that parents verbally ascribe implausibly sophisticated PAs to young children, who are reinforced when they respond in ways that can be followed up by still further

parental folk-psychological elaboration. By such processes children are incentivised to try to conform to the expectations encoded in the ascribed PAs, and to apply these PAs to themselves while they are forging their social identities. Children may come to be characterised by PAs that are generally understood by members of their communities not because they acquire these mental states through pre-progammed Piagetian development, with others then having to infer their presence, but because, in McGeer's phrase, the children are actively bootstrapped into a common practice of psychological interpretation.

A substantial list of theorists have gone on to independently generalise these suggestive observations into a model of *mindshaping*. This literature is consolidated by Zawidzki (2013), who then builds the most comprehensive elaboration and extension of the construct to a range of social and behavioural phenomena. As Zawidzki carefully argues, mindshaping need not displace mindreading altogether. If a person tries to influence someone's PAs and they resist, she might naturally be moved to hypothesise competing PAs they have already internalised and are unwilling to surrender. But the primary practical point of the mindshaping hypothesis is to address the fact that attempting to explain most social coordination by appeal to mindreading is empirically implausible.

The fundamental problem with asking too much of mindreading was identified by Morton (1996) in advance of the development of mindshaping theory. Folk psychology, the body of everyday default expectations about preferences and motivations, has little parametric structure; beliefs and desires can effectively be nested and combined ad infinitum. Thus if the point of mindreading were simply to rationalise behaviour, the task would not be demanding. Except in the case of people about whom one has detailed and relatively complete biographical information, which under assumptions of consistency builds up a restricted template of possible interpretations, one can always hypothesise a combination of beliefs and desires that would account for any observed behaviour. But this would amount to rampant curve-fitting, a disastrous strategy for making accurate out-of-sample predictions. This is a fatal problem given the primary point of the mindreading hypothesis, which is supposed to be that it explains everyday predictive successes that in turn explain social coordination.

When this sort of problem arises for theoretical inference in science, there are two generic kinds of response available: one can gather more data under deliberately controlled conditions, or one can add more structure to one's model specification. Where folk psychological inference is concerned, in the domain of day-to-day coordination with non-intimates the first response is ruled out: mindreading would be most needed in precisely those instances where the target has not had opportunities to furnish long runs of biographical narrative. People make judgments about one another's PAs rapidly and automatically, seldom conducting even simple tests. As for the alternative strategy, adding structure to hypotheses, this improves predictive power at the immediate cost of increasing expected error.

Mindshaping theorists emphasise a crucial respect in which social interaction is not like doing (prototypical) science: observation *normally* influences the phenomena. This leads quickly to the point that explanation is seldom among the goals of practical coordination; people aren't generally concerned with verifying or rejecting prior models of one another, but with achieving coordinated behaviour. Furthermore, except in cases of highly asymmetrical power or status, where order-giving is sufficient for coordination, social interaction is mutualistic: each interpreter can influence the other, and each may be willing to accede some normative authority for the sake of consensus.

It is a familiar observation that social interaction usually involves implicit bargaining, typically at fleeting and subtle real-time scales. Most people are consciously aware from time to time of engaging in such fine-scale bargaining. But the profoundly counter-Cartesian point of the mindshaping hypothesis is that folk psychology is a fundamentally *prescriptive* rather than descriptive structure (Morton 2003). It is not a *theory* but an *ideology* for quotidian collective social management.

Consider a simple imaginary example. Two strangers stand waiting on a subway platform. Suppose A remarks "The trains are a bit slow here." This *is* arguably a gamble on a weakly informed prediction that B will agree. If B does agree, the parties can enjoy fleeting solidarity against the transport authorities. But

suppose B replies "Oh, I don't know. I seldom wait longer than I mind." No socially aware human observer would think that A would need to pass through any true change of an underlying epistemic state if she then said "That's good to hear. I guess I'm just in a particular bit of a rush today." A socially conscientious B would know how to close the loop to complete concord by producing something along the lines of "I hear you! Work is such pressure these days."

Most people mildly enjoy exchanges of this kind. They are not at all pointless: the two parties have reinforced their feelings of being members of a common community because they have exchanged signals of their knowledge of how to quickly converge on shared PAs.[14] A major contributor to feelings of alienation among isolated immigrants is that they at least initially do not know how to achieve prosaic coordinations of this kind without serious effort and risk of embarrassment; the flip side is the fleeting but intense pleasure of going through the exercise with a chance-met co-national in such circumstances.

Quotidian mindshaping is the model for bargaining around PA ascriptions in less typical circumstances where more is at stake. An effective manager in an organisation is careful to minimise mindreading, but has learned how to mindshape in ways that preserve enough dignity and autonomy in others to foster team consciousness. Wise police officers do not try to *predict* what informants think; they try to induce them to frame their beliefs cooperatively. The dark side of mindshaping is that in situations of inter-group conflict it contributes to polarization: a person may signal solidarity with his own tribe by rejecting PAs an out-group member ascribes to him as a matter of general principle; by contrast, with a fellow insider he might acknowledge nuances in the opinions of the other tribe which, in the tension created by their presence, he would *sincerely* not notice.

Zawidzki and other mindshaping theorists do not deny the possibility or existence of mindreading, in the sense of explicit conjectures made 'on the fly' - for example, when an interactant appears to be evasive, makes implausible claims, or is caught in a conflict of interest. (Note that predicting another's strategic moves on the basis of careful analysis, which is simply due diligence in high-stakes interactions, is not mindreading on anyone's account; it is social research.) But it is extremely unlikely that this explains capacities for coordination; as Morton observed, and Zawidzki buttresses with extensive analysis and argument, mindreading is an ill-advised coordination strategy. Larrouy and Lecouteux (2018) produce a model demonstrating conditions under which mindshaping is a more effective device than mindreading for selecting Schelling-style focal points in formal coordination games; and the conditions in question are the standard, everyday ones.

In this context, Alós-Ferrer's (2018) suggestion that people are not 'really' social animals because their ape ancestors were more 'individualistic' than monkeys amounts to a perverse confusion of comparative scales *and* of dimensions. Regular engagement in mindshaping is the most intense form of sociality found in nature, because it involves organisms dynamically co-managing their behavioural control systems to cope in common with environments that rapidly change *because* of this special social facility. It is *possible* that humans are not the only mindshapers around, but if other animals do it the ranks do not appear to include our closest living relatives. (Realistic candidates are toothed whales, elephants, corvids, and parrots; see Ross [2019]. Note that behaviours selected for their influence on other individuals' beliefs and preferences, such as courtship dances or predator alarm calls, do *not* typically constitute mindshaping. Mindshaping is changing the patterns constructed by another individual when she assumes the intentional stance toward *herself*. Thus mindshaping presupposes socially conscious and explicit self-representation; hence its expected infrequency in nature.)

We therefore do not agree with Alós-Ferrer that economists can best contribute to social neuroscience by using game theory to model mindreading. The kind of strategic conjecture-and-test model that Alós-Ferrer

[14] It is worth stressing the externalist point again here: it would typically not be the case that either party started with a standing conviction about the efficiency of the subway system that the other set out to have revealed. PAs just are whatever is mutually negotiated dynamically.

evidently has in mind here reflects the style of game-theoretic modeling that, if applied to everyday social interactions, is diagnosed by Mirowski (2002) and Amadae (2016) as manifesting paranoia, the response of a Cartesian who *fears* that there are other minds. However, game theory, as a body of mathematics that can be supplemented with additional mathematical tools including network theory, is powerful and flexible. In the concluding section, we review its application to mindshaping. As mindshaping is something important and social that human brains evidently support, it offers a promising avenue indeed for contribution by economists to tri-disciplinary partnership with sociology and neuroscience.

6 Conditional game theory for modeling mindshaping

Over the course of about four decades to the mid-1980s, game theory gradually became the primary modeling technology of microeconomics. It achieved this status mainly because it allowed a maximally generalised concept of equilibrium (Nash equilibrium) to be applied at almost any scale of the modeler's choosing. This allowed economists to model competitive situations which are neither monopolistic nor perfect; and these are the situations actually confronted by most economic agents most of the time. One effect of the fusion of older microeconomic theory (essentially Pareto's refinement of that theory) with game theory was to lock into the axiomatic foundations of the discipline the identification of economic agents with consistent, acyclic, stable preference fields. Applied to people, this is of course a considerable idealisation, which is to say, a fiction. For the applications of most interest to economists, use of this fiction involves costs worth paying. Where outcomes of interactions are specifiable in terms of monetary prospects, or control of freely tradable assets with market-determined prices, people generally do have stable preferences for sustainably larger balance sheets over smaller ones. Even in these contexts, however, implicit negotiations around social status, or responses to perceived violations of fairness and other circumstantially sensitive norms, may matter to outcomes.

All of the above idealisations involve imposing boundaries on the flexibility of agency. Thus. one way in which they can be suspended is by *eliminating* agents from models. This is what evolutionary game theory does. In evolutionary games, strategies compete directly against one another for greater long-run frequency in populations. Without agents, there is no place for preferences or beliefs, limited in range or not.

Neither classical economic nor evolutionary game theory are well engineered for application to the problems that most typically interest sociologists (Luce and Raiffa 1957, p. 196). The units that sociologists study, people and institutions, take actions that reflect preferences and beliefs. One can abstract from this by building evolutionary models in which people and institutions are simply hosts for long-run competition among memes (Dennett 2016), but this is a highly misleading abstraction on the short scales where agents try to optimise their own degree of control over outcomes, not just the spread of the information they happen to have. Some 'behavioural' game theorists (Camerer 2003) have produced models in which preferences over distributions of social goods are simply inserted into the utility functions of agents. But this is ad hoc modeling (Binmore 2010), and in any event fails to engage with what most interests sociologists, who share the evolutionary game theorist's interest in the *dynamics* of transformation and reproduction of influence. We argued in the previous section that insofar as sociologists are motivated to take cognitive social phenomena into account, they should want to model mindshaping. But then both standard and evolutionary game theory look like incompletely developed tools for the job.

Recently, however, an innovation has been introduced into game theory, called Conditional Game Theory (CGT), by Stirling (2012, 2016), that is specifically designed to capture the dynamical propagation of preferences as conditioned on the strategic choices of individuals. This is precisely game theoretic representation of mindshaping, and CGT achieves this without hand-wiring social preferences into the modelled agents, thus avoiding Binmore's methodological criticism. The concepts that constitute solutions in CGT are unrefined Nash equilibria, so the whole accumulated analytical power inherited from the history of game

theory is preserved.

The distinction between standard game theory as an implementation of mindreading and conditional game theory as an implementation of mindshaping is best understood by way of contrast. Mindreading is a putative achievement of an individual reasoning by herself. Let z and z' be two alternatives for agent Z, let y and y' be two alternatives for agent Y, let $\succ$ denote an ordering mechanism "is preferred to", and suppose $z \succ z'$. Since Y can read Z's mind, Y knows that $z \succ z'$ and can use that knowledge to establish her preferences categorically with the ordering $y \succ y'$.

Mindshaping, by contrast, is an *interactive* phenomenon, whereby an individual exerts influence on others (in her role of mindshaper) and responds to the influence exerted on them (in the role of mindshapee) as they behaviorally negotiate their way to aligned respective desires and beliefs. "Aligned" does not necessarily mean "identical": consider agents of buyer and supplier firms, or two domestic partners, arriving over time at coordinated expectations about their respective obligations and entitlements. Zawidzki (2013) informally defines mindshaping as

> . . . a relation among four relata: a model, a target, a mechanism, and a set of respects in which the target can match the model. Mindshaping occurs when a mechanism aims to make a target match, in relevant respects, a model. The target is always the mind, that is, the categorical basis for some set of behavioral dispositions that characterize the agent. The mechanism can be some pattern of activity in an individual brain, as in basic forms of imitation, where the target's own neurally based mechanisms function to bring about a match between target and model [2013, p. 31].

Zawidzki's more precise specification is:

> . . . mechanism X mindshapes target Y to match model Z in relevant respects R, S, T, . . . , if and only if (1) effecting such matches is X's "proper function," . . . (2) X is performing its proper function, that is, causing Y to match Z in respect to R, S, T,. . . ; (3) Y is a mind, understood as a set of behavioral dispositions or the categorical basis for them; (4) X's performance of its proper function is guided by representations of R, S, T . . . ; and (5) Z is or is somehow derived from an agent other than the agent to which Y belongs [2013, p. 32].

Expressed in the register of preferences, let Z (the model, in Zawidzkian parlance) correspond to the preferences of the influencer, let Y (the target) correspond to the preferences of the influencee, and let X (the mechanism) correspond to the relation that causes Y to adjust (match) in response to Z. A graphical representation of this structure is $\mathrm{Z} \xrightarrow[\mathrm{X}]{} \mathrm{Y}$, meaning that Z influences Y via the mechanism X. Given a set of alternatives $\{z, z'\}$ for the influencer and a set of alternatives $\{y, y'\}$ for the influencee, consider the conditional statement "y is preferred to y' if z is preferred to z'", which corresponds to a hypothetical proposition with antecedent "z is preferred to z'" and consequent "y is preferred to y'". This scenario may be expressed symbolically as $y \succ y' \mid z \succ z'$, where "|" corresponds to the conjunction "if" with the antecedent on the right-side and the consequent on the left-side.

Y-the-mindreader infers that $z \succ z'$ by application of her cognitive powers, and need not even consider $z' \succ z$. Y's reasoning process may be completely private to her. By contrast, Y-the-mindshapee responds to an external social mechanism that connects her to Z-the-mindshaper. Any mechanism capable of distinguishing between Z's preference for z and Z's preference for z' must consider both $z \succ z'$ and $z' \succ z$. Thus, Y-the-mindshapee must define her consequent to both antecedents.

CGT implements mindshaping by applying the syntax of Bayesian probability theory to games so as to model conditional preferences as formally analogous to conditional probabilities. Individual players' preferences are modeled as conditional in two senses, at separate stages of analysis: as subject to influence

by other specific players, and as sensitive to the relative degrees of discord within groups that arises for different equilibria.

We sketch the essential features of the modeling, and begin by defining conjectures that players make about distributions of preferences. These should not be interpreted as attempts at mindreading; they are more accurately interpreted as models of a player wondering how her own preferences might turn out to strategically cohere, or not, with the distribution of other preferences in her society.

Let $\{X_1, \ldots, X_n\},\ n \geq 2$, represent a set of n players, and let $A_i = \{x_{i1}, \ldots, x_{iN_i}\}$ denote a finite set of actions available to X_i from which she must choose one element to instantiate, and let $\boldsymbol{A} = A_1 \times \cdots \times A_n$ denote the Cartesian product of the individual action sets. An action or strategy *profile* is an array: $\mathbf{a} \in \boldsymbol{A}$. In standard game theory, players have only 'categorical'—that is, unconditional—utility or payoff functions defined over strategy profiles: u_i: $\boldsymbol{A} \to \mathbb{R}$. We expand this by allowing for 'uncategorical' preferences that are conditional on preferences that others might currently have.

A *social influence network* comprises a directed graph with agents as vertices and influence relationships as edges. The expression$X_i \to X_j$ signifies that X_i (a *parent*) exerts social influence on X_j (a *child*). An agent is a *root vertex* if she has no parents. A *path*, denoted $X_i \mapsto X_k$, is a sequence of edges from X_i to X_k. A *closed path* is a path $X_i \mapsto X_i$. A directed graph is *acyclic* of there are no closed paths.

A *conjecture* $\mathbf{a}_i = (a_{i1}, \ldots, a_{in}) \in \boldsymbol{A}$ is a profile hypothesized by X_i as the outcome under consideration as the one to be instantiated. The element a_{ii} is X_i's *self-conjecture* and a_{ij}, $j \neq i$, is an *other-conjecture* by X_i for X_j. The array $(\mathbf{a}_1, \ldots, \mathbf{a}_n)$ is termed a *joint conjecture*.

Define the *parent set* $\mathrm{pa}\,(X_i) = \{X_{i_1}, \ldots, X_{i_{q_i}}\}$ as the subset of players whose preferences influence X_i's preferences. A *conditioning conjecture* by X_i for X_{i_k}, denoted $\mathbf{a}_{i_k} = (a_{i_k1}, \ldots, a_{i_kn})$, is a profile that X_i hypothesizes that X_{i_k} conjectures, $k = 1, \ldots, q_k$. The array $\boldsymbol{\alpha}_{\mathrm{pa}(i)} = (\mathbf{a}_{i_1}, \ldots, \mathbf{a}_{i_{q_i}}) \in \boldsymbol{A}^{q_i}$ is termed the *conditioning conjecture set*.

A *conditional utility function* defines X_i's preferences as conditioned on the conjectures of her parents: $u_{i|\mathrm{pa}(i)}(\cdot|\boldsymbol{\alpha}_{\mathrm{pa}(i)})$: $\boldsymbol{A} \to \mathbb{R}$. If $\mathrm{pa}\,(X_i) = \varnothing$ then the conditional utility $u_{i|\mathrm{pa}(i)}|\boldsymbol{\alpha}_{\mathrm{pa}(i)} = u_i$, the standard categorical utility. The collection $\{X_i, A_i, u_{i|\mathrm{pa}(i)}, i = 1, \ldots, n\}$ constitutes a finite, normal form, non-cooperative *conditional game*. The set $\{u_{i|\mathrm{pa}(i)}, i = 1, \ldots, n\}$ is the *utility framework*.

Through appropriate normalisation we can ensure that all utilities (i.e., categorical and conditional) are non-negative and sum to unity, which implies that the utilities have all of the characteristics of probability mass functions. If we restrict attention to networks that conform to two technical conditions, acyclicity and framing invariance (Stirling, 2012), then a conditional game satisfies the syntax of a Bayesian network—a directed acyclic graph with discrete random variables as vertices and conditional probability mass functions as edges. The fundamental theorem of Bayesian network theory (cf. Pearl 1988, Jensen 2001) is that the joint probability mass function is uniquely determined as the product of the conditional probability mass functions of all children's vertices and the unconditional probability mass functions of all root vertices. Applying this theory to conditional games, the analogue to the joint probability mass function is the *sociation model*, defined as

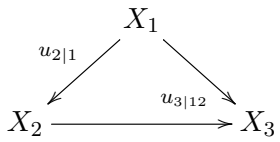
$$u_{1:n}(\mathbf{a}_1, \ldots, \mathbf{a}_n) = \prod_{i=1}^{n} u_{i|\mathrm{pa}(i)}(\mathbf{a}_i|\boldsymbol{\alpha}_{\mathrm{pa}(i)})\,. \tag{1}$$

The players'*ex post utilities* once social influence has permeated the group are determined by marginalization, yielding

$$u_i(\mathbf{a}_i) = \sum_{\sim \mathbf{a}_i} u_{1:n}(\mathbf{a}_1, \ldots, \mathbf{a}_n), \tag{2}$$

where $\sum_{\sim \mathbf{a}_i}$ means that the sum is taken over all arguments except $\mathbf{a}_i$. These ex post categorical utilities represent the players' preferences after taking into account the social relationships and interdependencies in the group. As these preferences are unconditional, standard solution concepts such as dominance and Nash equilibrium (NE) can be applied to them.

The technical condition of invariance implies that once the coordination utility has been defined, we can apply Bayes's rule to extract reciprocal influence relationships. Consider a three-agent conditional game with utility framework $\{u_1, u_{2|1}, u_{3|12}\}$ corresponding to the network

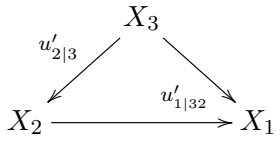


(3)

with sociation model

$$u_{123}(\mathbf{a}_1, \mathbf{a}_2, \mathbf{a}_3) = u_1(\mathbf{a}_1)u_{2|1}(\mathbf{a}_2|\mathbf{a}_1)u_{3|12}(\mathbf{a}_3|\mathbf{a}_1\mathbf{a}_2)\,. \tag{4}$$

This network may be reframed as $\{u'_{1|23}, u'_{2|3}, u'_3\}$, which corresponds to the network



(5)

with sociation model

$$u'_{123}(\mathbf{a}_3, \mathbf{a}_2, \mathbf{a}_1) = u'_{1|32}(\mathbf{a}_1|\mathbf{a}_3, \mathbf{a}_2)u'_{2|3}(\mathbf{a}_2|\mathbf{a}_3)u'_3(\mathbf{a}_3)\,, \tag{6}$$

where

$$u'_{1|32}(\mathbf{a}_1|\mathbf{a}_3, \mathbf{a}_2) = \frac{u_{123}(\mathbf{a}_1, \mathbf{a}_2, \mathbf{a}_3)}{u'_{32}(\mathbf{a}_3, \mathbf{a}_2)} \tag{7}$$

and

$$u'_{2|3}(\mathbf{a}_2|\mathbf{a}_3) = \frac{u'_{32}(\mathbf{a}_3, \mathbf{a}_2)}{u'_3(\mathbf{a}_3)} \tag{8}$$

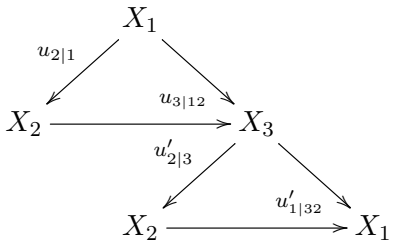
with

$$u'_{32}(\mathbf{a}_3, \mathbf{a}_2) = \sum_{\mathbf{a}_1} u_{123}(\mathbf{a}_1, \mathbf{a}_2, \mathbf{a}_3) \tag{9}$$

and

$$u'_3(\mathbf{a}_3) = \sum_{\mathbf{a}_1\,\mathbf{a}_2} u_{123}(\mathbf{a}_1, \mathbf{a}_2, \mathbf{a}_3)\,. \tag{10}$$

By direct substitution it is clear that (6) and (4) are equivalent, and that (5) is an inverse network for (3), namely,



(11)

In general, framing invariance means that different framings of a conditional game that use the same information, although encoded differently, yield the same sociation model. Reframing is an illustration of the power of mathematical models. Once the fundamental relationships are defined, the model can be manipulated in many ways to expose features that would otherwise be difficult to ascertain.

The sociation model provides an ordering over $\boldsymbol{A}$ for each conditioning conjecture set $\boldsymbol{\alpha}_{1:n} = (\mathbf{a}_1, \ldots, \mathbf{a}_n) \in \boldsymbol{A}^n$, and serves as a comprehensive model of all of the social relationships that exist among the individuals. However, since each agent is able to implement only its own self-conjecture, the critical issue for the

group is to consider behavior as a function of only the self-conjectures of the agents. Given a joint conjecture set $\boldsymbol{\alpha}_{1:n} = (\mathbf{a}_1, \ldots, \mathbf{a}_n)$, we form the *coordination profile*, $\mathbf{a} = (a_{11}, \ldots, a_{nn})$ and compute the marginal of the sociation model with respect to the coordination profile by summing over all elements of each $\mathbf{a}_i$ except the self-conjectures to form the *coordination function* for $\{X_1, \ldots, X_n\}$, yielding

$$w_{1:n}(a_{11}, \ldots, a_{nn}) = \sum_{\sim a_{11}} \cdots \sum_{\sim a_{nn}} u_{1:n}[(a_{11}, \ldots, a_{1n}), \ldots, (a_{n1}, \ldots, a_{nn})] \,. \tag{12}$$

Once the coordination function is defined, individual coordinated utilities can be extracted via marginalization, yielding

$$w_i(a_{ii}) = \sum_{\sim a_{ii}} w_{1:n}(a_{11}, \ldots, a_{nn}) \,. \tag{13}$$

We now consider a *tactical conditional game*. The conditional utility function $u_{i|\mathrm{pa}(i)}(\mathbf{a}_i|\boldsymbol{\alpha}_{\mathrm{pa}(i)})$ requires X_i to order her valuations for every conjecture $\mathbf{a}_i$ and for every joint conjecture set $\boldsymbol{\alpha}_{\mathrm{pa}(i)}$, which can quickly become arduous for even modestly complex networks. Fortunately, the introduction of conditional preferences creates the possibility of significant model simplifications and computational advantages that are not available with standard noncooperative game theory. The conditional utility $u_{i|j}(\mathbf{a}_i|\mathbf{a}_j)$ is the weight given to the statement "If X_j prefers outcome $\mathbf{a}_j$, then X_i prefers outcome $\mathbf{a}_i$". For many applications, however, a natural approach is for X_i to define her preferences over only her action set A_i given the conjectured actions of her parents, rather than defining her preferences over the entire outcome set $\boldsymbol{A}$, given the conjectured profiles of her parents. The utilities are then with respect to only the self-conjecture of X_i given the conjectured self-conjectures by X_i for her parents, yielding a *tactical conditional game* $\{X_i, A_i, \tilde{u}_{i|\mathrm{pa}(i)}, i = 1, \ldots, n\}$, where $\tilde{u}_{i|\mathrm{pa}(i)}(\cdot|\tilde{\boldsymbol{\alpha}}_{\mathrm{pa}\,(i)})$: $A_i \to [0, 1]$ with $\tilde{\boldsymbol{\alpha}}_{\mathrm{pa}\,(i)} = (a_{i_1 i_1}, \ldots, a_{i_{q_i} i_{q_i}}) \in \mathcal{A}_{i_1} \times \cdots \times \mathcal{A}_{i_q}$ denoting the set of conjectured self-conjectures by X_i for her parents. The tactical conditional utility $\tilde{u}_{i|j}(a_{ii}|a_{jj})$ is the weight given to the statement "If X_j prefers action a_{jj}, then X_i prefers action a_{ii}". The sociation model for a tactical conditional game is

$$w_{1:n}(a_{11}, \ldots, a_{nn}) = \prod_{i=1}^{n} \tilde{u}_{i|\mathrm{pa}(i)}(a_{ii}|\tilde{\boldsymbol{\alpha}}_{\mathrm{pa}\,(i)}) \,, \tag{14}$$

which thus becomes the coordination function. The coordinated utilities are given by (13).

We now extend conditional game theory to model networks with influence cycles. The fundamental theorem of Bayesian networks applies only to acyclic social influence relationships, which prohibits *independently specified* reciprocal relationships (i.e. Bayes's rule must be satisfied). In many social settings, however, social relationships are cyclic; that is, both $X_i \longrightarrow X_j$ and $X_j \longrightarrow X_i$, and are specified independently. Thus, it becomes necessary to extend from hierarchical network structures and accommodate cyclic network structures. In the interest of brevity and without loss of generality, we restrict attention to tactical conditional games. Consider the network

$$\begin{array}{ccc} X_1 & \xleftarrow{\tilde{u}_{1|k}} & X_k \\ {\scriptstyle \tilde{u}_{2|1}}\downarrow & & \uparrow \\ X_2 & \xrightarrow[\tilde{u}_{3|2}]{} & X_3 \end{array} \tag{15}$$

which may be expressed as a time-sequence of acyclic networks as follows. Let t denote the time required to traverse one cycle, let $\delta = 1/(k+1)$ denote the time required to traverse from X_i to X_{i+1}, let $X_i(s)$ denote X_i at time s, and consider the time-sequence

$$X_1(0) \xrightarrow[\tilde{u}_{2|1}]{} X_2(\delta) \xrightarrow[\tilde{u}_{3|2}]{} X_3(2\delta) \dashrightarrow X_k(k\delta) \xrightarrow[\tilde{u}_{1|k}]{} X_1(1) \cdots \tag{16}$$

which generates the path $X_1(0) \mapsto X_1(1)$. The coordination function corresponding to $X_1(0) \xrightarrow[\tilde{u}_{2|1}]{} X_2(\delta)$ is, following (14),

$$w_{12}(a_{11}, a_{22}, \delta) = \tilde{u}_1(a_{11})\tilde{u}_{2|1}(a_{22}|a_{11})\,, \tag{17}$$

from which the coordinated utility for X_2 at time δ is, following (13),

$$w_2(a_{22}, \delta) = \sum_{a_{11}} w_{12}(a_{11}, a_{22}, \delta)\,. \tag{18}$$

Continuing, the coordinated utility for X_3 at time 2δ is

$$w_3(a_{33}, 2\delta) = \sum_{a_{22}} \tilde{u}_{3|2}(a_{33}|a_{22}) w_2(a_{22}, \delta)\,. \tag{19}$$

In general,

$$w_{i+1}(a_{i+1i+1}, (i+1)\delta) = \sum_{a_{ii}} w_i(a_{ii}, i\delta)\tilde{u}_{i+1|i}(a_{i+1i+1}|a_{ii})\,. \tag{20}$$

Expressing this structure in matrix form, we define the *utility mass vector* at time $i\delta$ as

$$\mathbf{w}_i(i\delta) = \begin{bmatrix} w_i(x_{i1}, i\delta) \\ \vdots \\ w_i(x_{iN_i}, i\delta) \end{bmatrix}\,, \tag{21}$$

define the *agent-to-agent transition matrix*

$$T_{i+1|i} = \begin{bmatrix} \tilde{u}_{i+1|i}(x_{(i+1)1}|x_{i1}) & \cdots & \tilde{u}_{i+1|i}(x_{(i+1)1}|x_{iN_i}) \\ \vdots & \vdots & \vdots \\ \tilde{u}_{i+1|i}(x_{(i+1)N_{i+1}}|x_{i1}) & \cdots & \tilde{u}_{i+1|i}(x_{(i+1)N_{i+1}}|x_{iN_i}) \end{bmatrix}\,, \tag{22}$$

and it follows that

$$\mathbf{w}_{i+1}((i+1)\delta) = T_{i+1|i}\mathbf{w}_i(i\delta) \tag{23}$$

for $i = 1, \ldots, k$. Now define the *closed-loop transition matrix*

$$T_i = T_{i|i+k-1}T_{i+k-1|i+k-2}\cdots T_{i+2|i+1}T_{i+1|i}\,. \tag{24}$$

After t cycles,

$$\mathbf{w}_i(t) = T_i\mathbf{w}_i(t-1) = T_iT_i\mathbf{w}_i(t-2) = \cdots = T_i^t\mathbf{w}_i(0)\,. \tag{25}$$

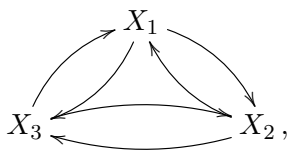
The key issue revolves around the convergence properties of T_i^t as $t \to \infty$. Under the appropriate technical restriction (regularity), the Markov convergence theorem (cf. Luenberger 1979) may be applied, which establishes that a) there exists a unique unity eigenvalue with corresponding normalized eigenvector $\overline{\mathbf{w}}_i$ of T_i such that $T_i\overline{\mathbf{w}}_i = \overline{\mathbf{w}}_i$; b) $\overline{T}_i = \lim_{t\to\infty} T_i^t = \begin{bmatrix}\overline{\mathbf{w}}_i & \cdots & \overline{\mathbf{w}}_i\end{bmatrix}$, and c) $\overline{\mathbf{w}}_i = \overline{T}_i\mathbf{w}_i(0)$ for every initial mass vector $\mathbf{w}_i(0)$. Thus, the coordinated utilities will converge to *coordinated steady-state utilities*,

$$\overline{\mathbf{w}}_i = \lim_{t\to\infty} \mathbf{w}_i(t) = \begin{bmatrix} \overline{w}_i(x_{i1}) \\ \vdots \\ \overline{w}_i(x_{iN_i}) \end{bmatrix}\,. \tag{26}$$

A feature of this result is that a closed-form solution exists for the converged individual utilities for all agents. This is a critical point: *It is not necessary for individuals to actually perform the iterations defined*

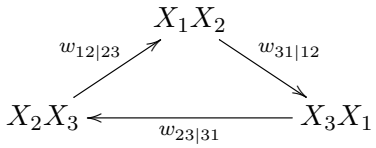
by (25). Assuming common knowledge, once the conditional utilities are specified and the social interchange is engaged, the individuals can immediately establish their coordinated preferences. (If knowledge is not common, then the agents will derive their own sociation models and arrive at socially "fuzzy" versions of coordinated behavior.)

To illustrate, consider the three-agent cyclic network



(27)

with $\mathcal{A}_i = \{x_{i i1}, x_{i2}\}$, $i \in \{1, 2, 3\}$, which consists of an outer cycle comprising all three agents and three inner cycles between adjacent agents. This structure suggests that we consider the relationships between pairs of agents and form the clockwise closed loop $\{X_1, X_2\} \to \{X_2, X_3\} \to \{X_3, X_1\} \to \{X_1 X_2\}$, namely,



(28)

where $w_{ij|jk}$ is the subnetwork-to-subnetwork influence function from $\{X_j, X_k\}$ to $\{X_i, X_j\}$. This function may be factored via the chain rule to obtain

$$w_{ij|jk}(a_{ii}, a_{jj}|a_{jj}, a_{kk}) = w_{j|ijk}(a_{jj}|a_{ii}, a_{jj}, a_{kk})\, w_{i|jk}(a_{ii}|a_{jj}, a_{kk}) \tag{29}$$

for $ij|jk \in \{12|23, 23|31, 31|12\}$, in which case $w_{i|jk}(a_{ii}|a_{jj}, a_{kk}) = \tilde{u}_{i|jk}(a_{ii}|a_{jj}, a_{kk})$. The function $w_{j|ijk}$, however, involves a *self-conditioning* component, since X_j is a member of both the influencer set $\{X_i, X_j\}$ and the influencee set $\{X_j, X_k\}$. Thus, $w_{j|ijk}$ is a degenerate mass function

$$w_{j|ijk}(a_{jj}|a_{ii}, a'_{jj}, a_{kk}) = \begin{cases} 1 & \text{if } a_{jj} = a'_{jj} \\ 0 & \text{otherwise} \end{cases} , \tag{30}$$

and (29) becomes

$$w_{ij|jk}(a_{ii}, a_{jj}|a'_{jj}, a_{kk}) = \begin{cases} \tilde{u}_{i|jk}(a_{ii}|a_{jj}, a_{kk}) & \text{if } a_{jj} = a'_{jj} \\ 0 & \text{otherwise.} \end{cases} \tag{31}$$

The subnetwork-to-subnetwork transitions are

$$\mathbf{w}_{ij} = T_{ij|jk} \mathbf{w}_{jk} \,, \tag{32}$$

where

$$T_{ij|jk} = \begin{bmatrix} w_{ij|jk}(x_{i1}, x_{j1}|x_{j1}, x_{k1}) & w_{ij|jk}(x_{i1}, x_{j1}|x_{j1}, x_{k2}) \\ w_{ij|jk}(x_{i1}, x_{j2}|x_{j1}, x_{k1}) & w_{ij|jk}(x_{i1}, x_{j2}|x_{j1}, x_{k2}) \\ w_{ij|jk}(x_{i2}, x_{j1}|x_{j1}, x_{k1}) & w_{ij|jk}(x_{i2}, x_{j1}|x_{j1}, x_{k2}) \\ w_{ij|jk}(x_{i2}, x_{j2}|x_{j1}, x_{k1}) & w_{ij|jk}(x_{i2}, x_{j2}|x_{j1}, x_{k2}) \\ w_{ij|jk}(x_{i1}, x_{j1}|x_{j2}, x_{k1}) & w_{ij|jk}(x_{i1}, x_{j1}|x_{j2}, x_{k2}) \\ w_{ij|jk}(x_{i1}, x_{j2}|x_{j2}, x_{k1}) & w_{ij|jk}(x_{i1}, x_{j2}|x_{j2}, x_{k2}) \\ w_{ij|jk}(x_{i2}, x_{j1}|x_{j2}, x_{k1}) & w_{ij|jk}(x_{i2}, x_{j1}|x_{j2}, x_{k2}) \\ w_{ij|jk}(x_{i2}, x_{j2}|x_{j2}, x_{k1}) & w_{ij|jk}(x_{i2}, x_{j2}|x_{j2}, x_{k2}) \end{bmatrix} \tag{33}$$

which, upon applying (31), becomes

$$T_{ij|jk} = \begin{bmatrix} \tilde{u}_{i|jk}(x_{i1}|x_{j1},x_{k1}) & \tilde{u}_{i|jk}(x_{i1}|x_{j1},x_{k2}) & 0 & 0 \\ 0 & 0 & \tilde{u}_{i|jk}(x_{i1}|x_{j2},x_{k1}) & \tilde{u}_{i|jk}(x_{i1}|x_{j2},x_{k2}) \\ \tilde{u}_{i|jk}(x_{i2}|x_{j1},x_{k1}) & \tilde{u}_{i|jk}(x_{i2}|x_{j1},x_{k2}) & 0 & 0 \\ 0 & 0 & \tilde{u}_{i|jk}(x_{i2}|x_{j2},x_{k1}) & \tilde{u}_{i|jk}(x_{i2}|x_{j2},x_{k2}) \end{bmatrix}. \tag{34}$$

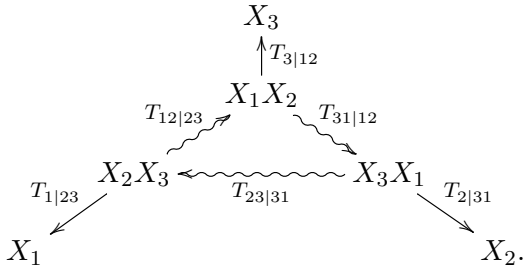
The steady-state subnetwork vectors, denoted

$$\overline{\mathbf{w}}_{ij} = \begin{bmatrix} \overline{w}_{ij}(x_{i1},x_{j1}) \\ \overline{w}_{ij}(x_{i1},x_{j2}) \\ \overline{w}_{ij}(x_{i2},x_{j1}) \\ \overline{w}_{ij}(x_{i2},x_{j2}) \end{bmatrix}, \tag{35}$$

are the eigenvectors corresponding to the unit eigenvalues of T_{ij}, where $T_{ij} = T_{ij|jk}T_{jk|ki}T_{ki|ij}$ for $ij|jk \in \{12|23, 23|31, 31|12\}$.

Once the network has converged, the influence relationships between the elements of the cyclic network are replaced by edges of the form $\rightsquigarrow$, indicating that the edges are *dormant*—they still exist but are inactive once steady state is achieved. Thus, the steady-state network becomes

(36)



The individual steady-state utility vectors are computed via

$$\overline{\mathbf{w}}_i = T_{i|jk}\overline{\mathbf{w}}_{jk} \tag{37}$$

where

$$T_{i|jk} = \begin{bmatrix} \tilde{u}_{i|jk}(x_{i1}|x_{j1},x_{k1}) & \tilde{u}_{i|jk}(x_{i1}|x_{j1},x_{k2}) & \tilde{u}_{i|jk}(x_{i1}|x_{j2},x_{k1}) & \tilde{u}_{i|jk}(x_{i1}|x_{j2},x_{k2}) \\ \tilde{u}_{i|jk}(x_{i2}|x_{j1},x_{k1}) & \tilde{u}_{i|jk}(x_{i2}|x_{j1},x_{k2}) & \tilde{u}_{i|jk}(x_{i2}|x_{j2},x_{k1}) & \tilde{u}_{i|jk}(x_{i2}|x_{j2},x_{k2}) \end{bmatrix}. \tag{38}$$

where the conditional utilities $\tilde{u}_{i|jk}$ for $i|jk \in \{1|23, 2|31, 3|12\}$ are specified by the problem statement.

Conditional game theory extends well beyond merely showing how to represent dynamics of social propagation of preferences. One possible motivation is aiding the design of distributed control architectures in autonomous agents that must, for the sake of efficiency, exploit both cooperative and competitive decision-making among sub-agents. There are important applications to the political economy of norm stabilisation and disruption, and to problems in welfare economics that can be addressed by identifying conditions in which an incentive-compatible group preference exists. (Where agents converge on a group preference in action we can 'fuse' the agents, thus allowing even for dynamics in the ontology of agents.) For present purposes, however, it suffices to show that the representation and modeling of mindshaping is not beyond the resources of game theory. For sociologists, one gloss on the general technology is that it provides a model of sociation (dissociation), the extent to which, in aggregate, agents are (aren't) sensitive to one another's preferences. Clearly, this is something that varies across historical societies and sub-societies. It plausibly has the status of a fundamental sociological variable.

And what of neuroscience here? Given the speed and fluidity with which people appear to engage in mindshaping exchanges, we can infer that their brains are prepared for the relevant learning. As imitation likely plays an important role, the prefrontal cortical areas that Alós-Ferrer (2018) identifies with a "Theory of Mind" network are plausibly implicated. There is not, however, current evidence for strong localization of social learning.[15] Furthermore, as both mindreading and mindshaping involve learning based on feedback from social interaction, they do not make clearly different predictions about supporting neural activity areas, at least given current knowledge of the functional pathways of the brain. Thus the recommended course of neuroscience research on mindshaping is not standard neuroeconomics based on neuroimaging data. What might be more interesting, and much in the spirit of the empirical approaches represented in Schutt et al (2015), would be to search for signs of chronic stress in people who recurrently encounter resistance to mindshaping they are obliged to attempt. One might, for example, study political canvassers who are working opposition-dominated neighbourhoods, or real estate agents during housing slumps, and compare them with random control subjects. Another approach might be to create experimental setups in which A-group subjects are asked to probe PAs in B-group subjects, but B-group subjects may not ask questions back. The mindreading hypothesis predicts asymmetries here, which might be reflected in behavioural measures of cognitive effort. As mindshaping can be conducted as readily in responder role as in questioner role, it doesn't so obviously predict similar asymmetries of effort. For analysis of any such experiments, an economist would want to recommend estimation of structural models rather than simple T-testing of null hypotheses. CGT provides the potential basis for such model specification and identification.

7 Conclusion

Because economics is fundamentally a social science, not a science of individual behaviour, neurosociology may offer a better context for interdisciplinary collaborations between economists and neuroscientists than the approaches that have featured to date in neuroeconomics, which borrows its hypotheses and experimental protocols from psychology. Economists have particular potential value to add in structurally modeling cognitive dimensions of sociality, thereby correcting for a current over-emphasis by sociologists on emotional responses. Neurosociologists are pursuing a broader and less problematic range of empirical methods than most neuroeconomists have done, and this provides helpful lessons to economists when hypotheses arising within their own discipline implicate mindshaping processes.[16]

References

Akerlof, G., and Kranton, R. (2010). *Identity Economics*. Princeton University Press.

Alós-Ferrer, C. (2018). A review essay on Social Neuroscience: Can research on the social brain and economics inform each other? *Journal of Economic Literature* 56: 234-264.

Amadae, S. (2016). *Prisoners of Reason*. Cambridge University Press.

Binmore, K. (2010). Social norms or social preferences? *Mind and Society* 2: 139-157.

Cacioppo, J., Visser, S., and Pickett, C., eds. (2006). *Social Neuroscience*. MIT Press.

Camerer, C. (2003). *Behavioral Game Theory*. Princeton University Press.

[15] Klucharev et al (2009) used fMRI to look for neural correlates of surprise and response adjustments when subjects were confronted with evidence that they were outliers relative to a reference sample in rating attractiveness of faces. They report, deeply unsurprisingly, activity in learning areas that are always observed, for well theorised reasons, when people encounter negative surprises. The 'conjunction analysis' they used to try to establish a control condition is subject to the standard econometric complaints about statistical inferences from fMRI studies.

[16] The authors thank Tad Zawidzki for his thoughtful comments on an earlier draft of this chapter.

Carruthers, P. (1996). *Theories of Theories of Mind*. Cambridge University Press.
Churchland. P. (1981). Eliminative materialism and the propositional attitudes. *Journal of Philosophy* 78: 67-90.
Churchland, P. (1988). *Matter and Consciousness*, Revised edition. MIT Press.
Clark, Andy (1997). *Being There*. MIT Press.
Clark, Andy (2003). *Natural Born Cyborgs*. Oxford University Press.
Clark, Austen (2001). Beliefs and desires incorporated. *Journal of Philosophy* 91: 404-425.
Damasio, A. (1994). *Descartes' Error*. Harper Collins.
Dekker, E., and Remic, B. (2019). Two types of ecological rationality: or how to best combine psychology and economics. *Journal of Economic Methodology* forthcoming.
Dennett, D. (1969). *Content and Consciousness*. Routledge.
Dennett, D. (1981). True believers. In A. F. Heath, ed., *Scientific Explanation*, pp. 150-167.
Dennett, D. (2017). *From Bacteria to Bach and Back*. Allen Lane.
Dreyfus, H. (1979). *What Computers Can't Do*. 2nd Edition. Harper and Row.
Fodor, J. (1975). *The Language of Thought*. Harvard University Press.
Frijters, P., and Foster, G. (2013). *Economic Theory of Greed, Love, Groups, and Networks*. Cambridge University Press.
Fumagalli, R. (2014). Neural findings and economic models: why brains have limited relevance for economics. *Philosophy of the Social Sciences* 44: 606-629.
Fumagalli, R. (2016). Five theses on neuroeconomics. *Journal of Economic Methodology* 23: 77-96.
Goldman, A, (2006). *Simulating Minds*. Oxford University Press.
Goyal, S (2007). *Connections*. Princeton University Press.
Graziano, M. (2013). *Consciousness and the Social Brain*. Oxford University Press.
Harrison, G. (2008). Neuroeconomics: A critical reconsideration. *Economics and Philosophy* 24: 303-344.
Herrmann-Pillath, C. (2013). *Foundations of Economic Evolution*. Edward Elgar.
Ioannides, Y. (2013). *From Neighborhoods to Nations: The Economics of Social Interactions*. Princeton University Press.
Jensen, F. (2001). *Baysian Networks and Decision Graphs*. Springer Verlag.
Klucharev, V., Hytönen, K., Rijpkema, M., Smidts, A., and Fernández, G. (2009). Reinforcement learning signal predicts social conformity. *Neuron* 61: 14-151.
Larrouy, L., and Lecouteux, G. (2018). Choosing in a large world: The role of focal points as a mindshaping device. HAL WP halshs-01923244. https://halshs.archives-ouvertes.fr/halshs-01923244
Luce, R., and Raiffa, H. (1957). *Games and Decisions*. Wiley.
Luenberger, D. (1979). *Introduction to Dynamic Systems*. Wiley.
Martens, B. (2004). *The Cognitive Mechanics of Economic Development and Institutional Change*. Routledge.
Martin, J. (2009). *Social Structures*. Princeton University Press.
McClamrock, R. (1995). *Existential Cognition*. University of Chicago Press.
McGeer, V. (2001). Psycho-practice, psycho-theory, and the contrastive case of autism. *Journal of Consciousness Studies* 8: 109-132.
Mirowski, P. (2002). *Machine Dreams*. Cambridge University Press.
Misyak, J.B., Melkonyan, T.A., Zeitoun, H., and Chater, N. (2014). Unwritten rules: virtual bargaining underpins social interaction, culture, and society. *Trends in Cognitive Sciences* 18: 512–519.
Morris, C. (1972). *The Discovery of the Individual: 1050-1200*. Harper and Row.
Morton, A. (1996). Folk psychology is not a predictive device. *Mind* 105: 119-137.
Morton, A. (2003). *The Importance of Being Understood*. Routledge.
Newell, A., and Simon, H. (1976). Computer science as empirical inquiry: Symbol and search. *Communications of the Association for Computing Machinery* 19: 113-126.

Nichols, S., and Stich, S. (2003). *Mindreading*. Oxford University Press.
Ofek, H. (2001). *Second Nature*. Cambridge University Press.
Pearl, J. (1988). *Probalistic Reasoning in Intelligent Systems*. Morgan Kaufman.
Pesonen, R. (2019). On the ecological and cognitive nature of mutual reasoning and decision making. This volume.
Pylyshyn, Z. (1984). *Computation and Cognition*. MIT Press.
Pylyshyn, Z., ed. (1987). *The Robot's Dilemma: The Frame Problem in Artificial Intelligence*. Ablex.
Rizzolatti, G., Fogassi, L., and Gallese, V. (2001). Neurophysiological mechanisms underlying the understanding and imitation of action. *Nature Neuroscience Reviews* 2: 661-670.
Ross, D. (2005). *Economic Theory and Cognitive Science: Microexplanation*. MIT Press.
Ross, D. (2008). Two styles of neuroeconomics. *Economics and Philosophy* 24: 473-483.
Ross, D. (2013). The evolution of individualistic norms. In K. Sterelny, R. Joyce, B. Calcott, and B. Fraser, eds., *Cooperation and its Evolution*, pp. 17-43. MIT Press.
Ross , D. (2014). *Philosophy of Economics*. Palgrave Macmillan.
Ross, D. (2019). Consciousness, language, and the possibility of non-human personhood: Reflections on elephants. *Journal of Consciousness Studies* 26: 227-251.
Schutt, R., Seidman, L., and Keshavan, M., eds., (2015). *Social Neuroscience*. Harvard University Press.
Smith, V. (2008). *Rationality in Economics*. Cambridge University Press.
Sterelny, K. (2003). *Thought in a Hostile World*. Blackwell.
Sterelny, K. (2012). *The Evolved Apprentice*. MIT Press.
Stirling, W. (2012). *Conditional Game Theory*. Cambridge University Press.
Stirling, W. (2016). *Theory of Social Choice on Networks*. Cambridge University Press.
Taylor, C. (1989). *The Sources of the Self*. Harvard University Press.
Turner, J. (2015). The neurology of human nature. In R. Schutt, L. Seidman, and M. Keshavan, eds., *Social Neuroscience*, pp. 41-87. Harvard University Press.
Wellman, H. (1990). *The Child's Theory of Mind*. MIT Press.
Zawidzki, T. (2013). *Mindshaping*. MIT Press.