PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B

Phil. Trans. R. Soc. B (2012) **367**, 2224–2233
doi:10.1098/rstb.2012.0124

Review

The economic approach to 'theory of mind'

Nikolaus Robalino and Arthur Robson*

Department of Economics, Simon Fraser University, Burnaby, British Columbia, Canada

Theory of mind (*ToM*) is a great evolutionary achievement. It is a special intelligence that can assess not only one's own desires and beliefs, but also those of others. Whether it is uniquely human or not is controversial, but it is clear that humans are, at least, significantly better at *ToM* than any other animal. Economists and game theorists have developed sophisticated and powerful models of *ToM* and we provide a detailed summary of this here. This economic *ToM* entails a hierarchy of beliefs. I know my preferences, and I have beliefs (a probabilistic distribution) about your preferences, beliefs about your beliefs about my preferences, and so on. We then contrast this economic *ToM* with the theoretical approaches of neuroscience and with empirical data in general. Although this economic view provides a benchmark and makes useful suggestions about empirical tendencies, it does not always generate a close fit with the data. This provides an opportunity for a synergistic interdisciplinary production of a falsifiable theory of bounded rationality. In particular, a *ToM* that is founded on evolutionary biology might well be sufficiently structured to have predictive power, while remaining quite general. We sketch two papers that represent preliminary steps in this direction.

Keywords: theory of mind; evolution; game theory; economics

1. INTRODUCTION

Human cognition, the theme of the present collection, is clearly an evolutionary achievement of the highest order. A capstone attribute of human cognition is theory of mind (*ToM*), and this is the topic of the current contribution (and a major focus of Frith's contribution in this issue [1]). An agent with *ToM* has the ability to conceive of himself, and of others, as intentional beings. At the core of this is an ability to attribute mental states—desire, knowledge, belief, intent, etc.—and to interpret observed behaviour in terms of such states. Although the phrase '*ToM*' rarely appears in the economics literature, the foundations of modern economic theory include a powerful and elegant mathematical description of *ToM*. This theory is germane to decision theory, and is crucial to the study of strategic interaction.

The central economic assumption here about behaviour is that individuals are Bayesian rational. In strategic settings, this assumption needs to be supplemented, most fundamentally by the assumption that there is 'common knowledge' of such Bayesian rationality. An individual who is Bayesian rational is endowed with well-defined preferences and subjective probabilistic beliefs about the world and makes optimal choices in the light of these preferences and beliefs, where these optimal choices are generally taken as those implied by maximization of expected utility. Further, such an agent updates her beliefs in the light of new information in the mathematically correct fashion described by Bayes' theorem. Common knowledge of rationality (CKR) means that each individual knows that each other individual is rational; that each individual knows that each other individual knows that the first individual is rational; and so on, *ad infinitum*.

Theoretical research in economics has largely concerned itself with the sometimes surprisingly subtle implications of Bayesian rationality for decision theory and game theory. This theory generally progressed with minimal input from experimental data. Although this stance might seem unreasonable, it has been productive until now. (The theory of auctions is an example of a success story within economic theory. Perhaps because auctions are often high-stakes games, quite subtle theoretical effects can sometimes be found in the observational data. Klemperer [2], for example, surveys auction theory.) The economic stance is best understood as an agnosticism about internal mental processes that focuses instead on idealized observed behaviour.

Recently, work in experimental and behavioural economics (importantly inspired by psychology) has recorded systematic deviations from the Bayesian rational ideal. Early work focused on the choices of individuals in isolation. Later work tested behaviour in the context of strategic or social interaction problems. The work here suggests important cognitive limitations that manifest themselves in behaviour that is not consistent with the standard model. Two of the most salient departures are associated with social preferences, such as altruistic or spiteful behaviour, and computational limits. The time now seems ripe, then, to initiate the synergistic integration of these two approaches—the top-down theoretically driven approach of economics and the bottom-up more empirically motivated approach of neuroscience and psychology.

*Author for correspondence (robson@sfu.ca).

One contribution of 15 to a Theme Issue 'New thinking: the evolution of human cognition'.

In this spirit, the present purpose is to provide a short and informal survey of the traditional economic approach to full sophistication in strategic interactions, and to consider how this approach should be modified in the light of neuroscientific theory and empirical observation. In greater detail, we set the stage by presenting a sketch of the economic theory of own preferences in a non-strategic setting, as well as a derivation of such preferences from a biological model. We next sketch the theory of fully sophisticated strategic interactions that rely upon CKR. We then discuss some of the implications of neuroscience and experimental studies in relation to how people deviate from the rational ideal. We conclude by considering a biological model in which it would be advantageous for an individual to represent the preferences of other agents. This last model extends in a natural fashion the model of the evolution of own preferences.

2. UTILITY AND PREFERENCES IN ECONOMICS

The term 'rationality' has a rather weak meaning in economics. If individuals maximize *some* utility function, such utility functions must satisfy only very weak restrictions. Indeed, the standard approach to demand theory is via 'revealed preference' [3] (see ch. 1–3 in [4] for a modern textbook treatment). Revealed preference investigates the restrictions on choice behaviour in the relevant market settings that are implied by utility maximization. In particular, it derives a set of restrictions that is both necessary and sufficient for the derivation of the entire class of associated utility functions. (This class of utility functions have the same indifference curves, roughly speaking.) This set of restrictions on choice behaviour represents the culmination of a historical process of applying 'Occam's razor' to strip away unnecessary assumptions on utility functions.

Why do economists assume that individuals are Bayesian rational, in particular? It is not that economists believe that each individual consciously and unerringly maximizes expected utility, where the expectation is derived from her beliefs and where these beliefs are manipulated precisely as prescribed by Bayes theorem. In the first place, there is no need for conscious deliberation. Binmore ([5], p. 61) provides a nice account of the standard analogy—

> In keeping their balance, (cyclists) do not solve complex systems of differential equations in their heads (or, at least, not consciously). Nor are they conscious of monitoring such factors as windspeed and the camber of the road. Nevertheless, they behave as though they were consciously gathering all the necessary data and then computing an optimal response.

That is, the standard economic philosophy is that it is 'as if' individuals were Bayesian rational and it is not necessary that this be an accurate description of the *process* by which this is achieved.

In the second place, it is not necessary to believe that this model describes any individual's behaviour exactly. As is inevitable in any account of human behaviour, we hope that the model captures only a central tendency.

However, sufficiently egregious non-compliance with some aspects of Bayesian rationality might be selected against. Such selection might be social in nature—it might involve being taken advantage of by others as in a 'money pump', for example. Consider, that is, an individual whose preferences are intransitive, so that he has the following strict preferences over three bundles of goods A, B and C as follows: $A \succ B \succ C \succ A$. Indeed, given a binary choice, he would prefer the 'better' option even if he has to also pay \$1. Suppose he is endowed with the bundle A. Now he is offered a chance to swap A for C, which he does, also conceding \$1. He is then offered a chance to swap C for B, which he does, again giving up \$1. Finally, he is offered back the original bundle A, which he accepts, giving up a further \$1. He is now back where he started, except for being out of pocket to the tune of \$3. This process can then be repeated arbitrarily often, indeed pumping away all the individual's wealth. More generally, but less rapidly, biological evolution might serve to weed out individuals whose behaviour departs radically from the Bayesian ideal. We consider later a simple example of this.

Let us turn now to perhaps the central construction in decision theory—the derivation of the 'expected utility' criterion for choice under uncertainty. We begin with the expected utility theorem of von Neumann & Morgenstern [6]. In contrast to the idiosyncratic subjective probabilities considered by Savage [7], which may vary arbitrarily between individuals, von Neumann & Morgenstern consider the simpler case of objective probabilities, probabilities for which there is general agreement. The basic approach in both cases is axiomatic—various general underlying principles to guide choice are suggested, where these axioms have intrinsic intuitive appeal. In the von Neumann & Morgenstern case, these axioms generate a criterion based on the expectation of the utility of each possible outcome.

The crowning expression of Bayesian rationality in decision theory is the axiomatic derivation by Savage [7] of expected utility with subjective probabilities. That is, these probabilities are no longer objective, agreed upon, as in von Neumann & Morgenstern, but are idiosyncratic, part of preferences themselves. Savage imposes all the axioms used by von Neumann & Morgenstern that ensure the requisite additive separability in the subjective probabilities. He must also impose additional axioms to identify these subjective probabilities from behavioural choices.

Savage's treatment of expected utility supposes that an agent considers all possible states of the world in formulating her preferences. The notion of 'states of world' has been refurbished in modern epistemological game theory, where states of the world may now include the strategies chosen by opponents or by the opponents' 'types'. In a two person game, my opponent's type may include her payoffs, her beliefs about my payoffs, her beliefs about my beliefs about her payoffs and so on. We return to consideration of modern game theory in §3. First, we conclude the present section on decision theory by considering the biological evolution of expected utility.

(a) *The evolution of utility*

What light does biological evolution shed upon preferences; or on attitudes to risk in particular? Perhaps the best metaphor for thinking about the relationship of evolution to the individual comes from economics—the principal–agent problem. Consider a firm with one owner (one for simplicity) and one CEO. The owner wishes the CEO to maximize the value of the firm, but the owner lacks the detailed information about the circumstances of the firm that is readily available to the CEO. The owner would then like to design a contract for the CEO that ensures that the CEO is induced to come as close as possible to maximizing this value of the firm.

In the same way, evolution might be thought of as a principal who 'wishes' the individual as the agent to maximize reproductive success. However, still as in the principal–agent metaphor, evolution is not aware of various idiosyncratic features of the environment. That is, to use less anthropomorphic and more neutral language, these features were not frequent enough in evolutionary history that behaviour appropriate to the feature was selected. Robson [8] argues that such a situation favours the partial devolution of control away from evolution onto the individual. That is, it pays for evolution to allow the individual to freely decide on an action, while still dictating the individual's preferences over some appropriately specified outcomes. This combines the benefits of setting the right goals with exploiting those arising from the local information of the individual.

Such an argument is applied in Robson [8] to explain the evolution of attitudes to risk, in particular. Suppose that there a finite set of outcomes, $c_1, \ldots, c_N$ say, each of which has an associated level of fitness. Suppose we hypothesize that the individual has been equipped by evolution with a utility for each outcome that agrees with fitness. Arbitrary gambles arise over these outcomes, and the individual should favour the gamble with the higher level of expected utility. There are infinitely many more such gambles, of course, than there are outcomes. The individual has to choose between two such gambles, but she does not know the probability distribution of either gamble to begin with. Rather she has a prior belief distribution over each distribution. However, she has many opportunities to make a draw from one or the other gamble and can update her belief distribution for the gamble chosen.

This problem is the so-called two-arm bandit problem, by analogy with the 'one-armed bandit' found in casinos. In general, this is a problem that is made difficult by the tension that exists between the desire to exploit the arm that has the higher apparent expected payoff, and the desirability of experimentation to ascertain more clearly which arm is preferable in fact. That is, you might be tempted to choose the arm that you believe is better currently, but this choice means that you give up on checking whether the other arm, despite seeming worse now, might nevertheless turn out to be preferable. The two-armed bandit problem, although trivial in principle because there are only a finite number of options overall, is famously difficult to analyse from a practical point of view.

Consider the dramatic simplification, however, that arises if there are many repetitions, with each one having negligible fitness consequences. In this case, experimentation is cheap and it is possible to have one's cake and eat it too. That is, it is possible to ascertain which arm is preferable using only a small fraction of the repetitions, and then to exploit that arm for the preponderance of the time.

What we have argued so far is that, if an individual is equipped by evolution with a utility of each outcome that agrees with fitness, then she can learn to choose correctly between an arbitrary unknown pair of distributions on the two arms. That is, she can react appropriately to entirely novel distributions, as long as she had been exposed to a set of previous distributions that was sufficiently rich to tie down the utility for all outcomes. The possession of a utility function allows evolutionarily optimal behaviour to be generated in a decentralized fashion by the individual. Indeed, any method of ensuring that the individual can adapt to such novel distributions must have implicit in it the same utilities over outcomes, and entail the maximization of expected utility.

3. BAYESIAN THEORY OF MIND

While decision theory models rational choice in isolation, game theory provides a description of how rational actors interact with one another. Its formal apparatus has yielded a convenient framework for modelling a great number of social problems—broadly described as issues of conflict and cooperation. Economists have employed these methods to study a seemingly limitless variety of social and economic matters, e.g. the hiring decisions of firms, the bargaining behaviour of employers and unions, the designing of contracts, the voting decisions of a population, and even courting and mating behaviour.

In strategic choice, the consequences of an act will, in general, depend on the choices made by other agents. Hence, in addition to being able to identify a favourite among a set of outcomes, an actor must be able to predict the behaviour of others. In order to do this, he might have to reflect on the desires (or utility functions) of his opponents. It could, however, be premature for his deliberations to stop there. A natural consequence of strategic thinking is an internal dialogue akin to the following.

> D: What she does depends on what she thinks I will do, but what she thinks I will do depends on what she thinks I think she will do, but what she thinks I think she will do . . .

In order to settle the question of what economic agents will do, it appears that we, as game theorists, require a complete, explicit description of their internal states of mind—these internal states including, crucially, beliefs about the mental states of their opponents. On the face of it, we need a theory of *ToM*, a way to represent sequences of nested, self-referential descriptions of internal states, e.g. the first player's beliefs about the second player's beliefs about the first player's intents. Although in many applications of game theory the earlier-mentioned infinite regress is circumvented by

the notion of a Nash equilibrium [9,10] (in a Nash equilibrium, no player can improve her payoff by unilaterally deviating from her equilibrium strategy), game theorists have still found it necessary, in order to analyse some situations, to develop formal models of full-fledged theories of mind.

Most of this effort has been in two branches of the literature. The first, initiated by Harsanyi [11,12], was motivated by the analysis of games in which players lack information about substantive details of the game. Harsanyi recognized that a complete description of such a game required the explicit representation of an infinite regress of reciprocal beliefs. He arrived at this by reasoning in the following way. The introduction of a parameter that is unknown to some player necessitates a description of his probabilistic beliefs about the unknown parameter. Suppose these beliefs, themselves, are unknown to the other players. Then an account of other actors' probabilistic beliefs about the first player's beliefs must be given. This in turn requires an account of all players' beliefs about those beliefs, and then a description of beliefs about those beliefs and so on. Even in simple environments, an exhaustive description could result in a staggeringly complex, intractable structure. Harsanyi provided a tractable way to analyse such problems. His approach was to recognize that the entire hierarchical belief structure could be summarized by modelling each player's mental configuration as a utility function paired with his probabilistic beliefs about the other players' mental configurations [13]. Nested in a player's beliefs about others' mental configurations, there are beliefs about the first player's mental configuration, and so on. Thus, the Harsanyi construction is an explicit *ToM*.

The second branch of the literature to give a serious account of players' internal states of mind aims to provide epistemic foundations for the equilibrium concepts themselves. A central question in this area is: what do players need to know (about a game, about others and about what others know) for their actions to constitute an equilibrium? Possibly the best known, certainly the most powerful, answer to the this question is given by Aumann [14], who shows that common knowledge of Bayesian rationality induces correlated equilibria, a generalization of Nash equilibrium allowing correlation in strategies. Here, common knowledge, by Alice and Bob, of an event means both Alice and Bob know it, Alice knows that Bob knows it, Bob knows that Alice knows it, Alice knows that Bob knows that Alice knows it and so on (the idea was formalized in [15]; for surveys of the concept, and particularly its use in economics, see [16,17]).

Harsanyi's model of *ToM* admits a succinct description of a potentially intractable hierarchy of beliefs that is amenable to equilibrium analysis. His construction of, say, Alice's mental configuration is composed of: (i) probabilistic beliefs about her own utility function, and (ii) a probability distribution over the possible mental configurations of Bob. Note that embedded in the mental configurations of Bob that Alice deems possible are hypothetical mental configurations of Alice—the mental models of Alice that Alice believes Bob deems possible. In fact, within each mental model, there is an infinite sequence of self-referential

comedy

Bob \ Alice	go	stay
go	2, 2	−3, 0
stay	0, 2	0, 0

other

Bob \ Alice	go	stay
go	2, −3	−3, 0
stay	0, −3	0, 0

Figure 1. Payoffs when the featured film is a comedy and when it is not a comedy.

models. Still, even though the Harsanyi model describes a situation with an infinite number of nested *ToM*s, his compact formulation permits a complete analysis based on only the first layer, a player's probabilistic beliefs about own utility and his beliefs about the minds of other players. The remainder of this section is dedicated to clarifying how this works.

(a) *A motivating example: Alice and Bob go to the movies*

We will illustrate the main ideas in a game theoretic context using the story of two friends, Alice and Bob.

There is one movie theatre in town and it features a film at 7 pm. The friends must choose, in ignorance of each others' choices, whether to go to the movies or stay at home. Of all types of films, Alice only enjoys only comedies and has a distaste for any other type of movie. Bob, on the other hand, is somewhat indifferent between all movies. In fact, he can enjoy only the movie experience if Alice is there with him and has a decidedly negative experience when he watches a film without her. Figure 1 describes the game payoffs.

Of the two numbers in each cell, the first is the payoff corresponding to Bob, and the second is Alice's payoff. For example, if Alice goes to the theatre and Bob stays at home, and the film is a comedy, the payoffs are given by the lower-left cell of the first matrix. Bob gets zero and Alice obtains a 2. On the other hand, if the movie is not a comedy and both friends go to the theatre, Bob is happy to watch a film with his friend. He gets 2. Alice gets -3.

Figure 1 describes the friends' material payoffs, but more is needed to settle the matter of how we should expect them to play the game—this will depend on what they know about each others' state of knowledge.

As a first pass, suppose Alice knows the type of film being featured. In the event the feature is a comedy, she is better off going to the theatre *irrespective* of what Bob does. Alternatively, if the film is not a comedy, she is better off staying home—again irrespective of what Bob does. Hence, if Alice is rational and knows the genre of the featured film, her strategy should be: go to the theatre if the movie is a comedy, and stay at home otherwise.

In order for Bob to choose appropriately, he must predict how Alice will behave. To do this, he should know the genre of the film and, moreover, that Alice knows it too. Then, by reasoning as we did previously, he can conclude that she will go to the theatre when a comedy is screened and stay at home otherwise. His rational choice is then to go to the theatre when the movie is a comedy and stay at home when it is not.

feature:	film 1	film 2	film 3
Alice knows	1 or 2	1 or 2	3
Bob knows	1	2 or 3	2 or 3

Figure 2. The information of the two friends.

We have just described the equilibrium of the game (the friends go to the theatre when the film is a comedy and stay at home when it is not) when it is assumed that Alice and Bob know the genre and that Bob knows that Alice knows it. It should come as no surprise that we obtain a different prediction when we change these assumptions.

(b) *Alice and Bob with incomplete information*

Now suppose the friends do not know the genre of the film being featured at 7 pm. We proceed to model the epistemological environment using an informational model developed in [7,15]. It will later be shown that this information model is associated with an equivalent Harsanyi structure in that it induces the same hierarchy of beliefs.

Let us assume Alice and Bob, at the moment of deciding whether or not to go to the theatre, have different vantage points of the marquee. Suppose further that there are three films that could potentially be screened at the local theatre, movies 1, 2 and 3, *Blazing Saddles*, *The Godfather* and *The Odd Couple*, respectively. Alice and Bob know this. Moreover, they know 1 and 3 are comedies and that 2 is not. However, Alice is far away from the marquee when she decides whether to go to the theatre or not. From where she sits she cannot read the name of the feature, but can only count the words in the title. Bob, on the other hand, lives close enough to the matinée, but is situated at a disadvantageous angle with respect to the marquee. A large building covers his view of all but the first three letters of the advertisement. When the theatre advertises that it will showcase *Blazing Saddles*, Bob knows a comedy is being screened but Alice cannot distinguish between the announcements, 'screening *Blazing Saddles*' and 'screening *The Godfather*'. When *The Godfather* is advertised, both of the friends are unsure of the feature's genre. Alice knows the screening is either of *Blazing Saddles* or of *The Godfather* while Bob knows it is either *The Godfather* or *The Odd Couple*. Finally, when *The Odd Couple* is scheduled, Alice knows what is being screened but Bob is unsure of whether the actual feature is *The Godfather* or *The Odd Couple*. Figure 2 describes the friends' vantage points.

When the friends are Bayesians, it is possible to calculate their posterior (after assuming their vantage points) beliefs about the genre of the film using their prior (before assuming their vantage points) beliefs about the film. For specificity, suppose before assuming their vantage points, the two friends believe film 1, film 2 and film 3 are all equally likely. Then, for example, when the state is 1 and Alice observes the marquee, she knows that the state is either 1 or 2. Because these mutually exclusive events are equally likely to occur, she must, as a Bayesian, assign probability $\frac{1}{2}$ to the feature being film 1 and $\frac{1}{2}$ to the feature being 2. Proceeding in this manner, we can obtain the friends' posterior beliefs about the type of film. These are displayed in figure 3.

feature:		film 1	film 2	film 3
Alice's beliefs:	comedy	1/2	1/2	0
	other	1/2	1/2	1
Bob's beliefs:	comedy	1	1/2	1/2
	other	0	1/2	1/2

Figure 3. Beliefs after assuming their vantage points when Alice and Bob believe each of the films is equally likely—before assuming their vantage points.

(c) *Generating mind theories*

We have thus far provided a description of how the friends see the world with regards to the underlying state of uncertainty—the genre of the featured film. But suppose Alice reasons that Bob must have a mind just like she does. In her deliberations about his mind, she might come to the conclusion that he maintains some notion of how she herself sees the world.

To see how we can generate a *ToM* from the information structure from figure 2, assume now that Alice believes figure 2, her own, and Bob's prior beliefs are common knowledge, and moreover that she believes the event 'Alice and Bob are Bayesians' is common knowledge. Consider the event $E =$ 'the feature is a comedy' which occurs whenever film 1 or film 3 are scheduled. If the event 1 is advertised, Alice knows that the feature is film 1 with probability $\frac{1}{2}$ and film 2 with probability $\frac{1}{2}$. She knows 1 is a comedy. Therefore, she assigns probability $\frac{1}{2}$ to event E whenever film 1 is scheduled. These are her first-order beliefs. Since she knows figure 3, Alice knows that whenever the actual feature is 1, Bob will know it. Therefore, she assigns probability $\frac{1}{2}$ to the event 'Bob knows the film is a comedy'. She also knows that when the screening is of 2, Bob assigns probability $\frac{1}{2}$ to it being a comedy. Thus, she assigns the residual probability of $\frac{1}{2}$ to the event 'Bob assigns probability $\frac{1}{2}$ to the feature being a comedy'. Those are Alice's second-order beliefs (beliefs about Bob's first order beliefs), conditional on film 1 being advertised. In a similar way, we can calculate her third-order beliefs—her beliefs about Bob's beliefs about her first-order beliefs, and so on.

(d) *Harsanyi's type structure*

Recall now our informal description of Harsanyi's *ToM* model. For example, Alice's mental configuration is a probability distribution over possible pairings of: (i) a utility function for herself and (ii) a mental configuration for Bob. In this section, we construct Harsanyi's type structure for the environment described in figure 2. In accordance with standard game theory terminology, from now on we refer to a player's mental configuration as his or her type.

For an intuition for how to generate the type space, consider the following. For any given underlying state of the world (a scheduled film), a player either knows or does not knows the film. But for each player, not-knowing can happen in only one way. When Alice

feature:	film 1	film 2	film 3
Alice knows	t_A^2	t_A^2	t_A^1
Bob knows	t_B^1	t_B^2	t_B^2

Figure 4. Mapping from features to types. The '1' types know the feature. The '2' types know the feature is one of two films.

does not know the feature she knows that the feature is either 1 or 2. Moreover, whenever a player knows the feature, he/she believes the other player is the not-knowing type. On the other hand, whenever a player is a not-knowing type, he/she assigns probability $\frac{1}{2}$ to the other player being the knowing type and $\frac{1}{2}$ to the other player being the not-knowing type. Thus, we can generate a type structure for our model with four types, denoted t_A^1, t_A^2, t_B^1 and t_B^2. The types with the subscript A are Alice's; those with the subscript B are Bob's. The two different types for each player correspond to situations in which they either know the film being screened or they do not know it.

When the feature is 3, Alice is type t_A^1, which knows the film is 3 and assigns probability 1 to Bob being type t_B^2 (the not-knowing type). When the feature is either 1 or 2, Alice is type t_A^2 which assigns probability $\frac{1}{2}$ each to the events 'the feature is 1 and Bob is type t_B^1' and 'the feature is 2 and Bob is type t_B^2'. Bob's types can similarly be defined (figure 4).

To see that this type structure will generate precisely the same beliefs as those obtainable directly from figure 2, consider Alice's beliefs about the event $E = \{$'the feature is a comedy'$\}$ when 1 is scheduled. In this case, Alice is type t_A^2 and Bob is t_B^1. Reading the t_A^2 beliefs from the top table of figure 5, we see that Alice assigns probability $\frac{1}{2}$ to event E. So her first-order beliefs about E are the same as those obtained in §3*c*. Furthermore, she assigns probability $\frac{1}{2}$ to each of the events 'Bob is t_B^1' and 'Bob is t_B^2'. But t_B^1 assigns probability 1 to E, and t_B^2 assigns probability $\frac{1}{2}$ to E. Clearly, Alice has precisely the same second-order beliefs about E as those we obtained in the partitional model. In fact, the beliefs coincide for all events, in all states.

(e) *What have we done so far?*

Thus far, we have described two compact representations of a hierarchical *ToM* (figures 2 and 5). The reader might be asking: what is the point of this, aside from providing a convenient short-hand? The answer is that the compact representations admit tractable equilibrium analyses.

The traditional sentiment among game theorists is that in incomplete information environments, it is the hierarchies, rather than type structures etc., that really underlie strategic interaction. Analysing games through belief hierarchies—which can obtain boundless depths—is often impracticable, however. The power of Harsanyi's approach lies in the fact that reasonable behaviour in the real game, in which players act under maintained hierarchies of belief, is also a reasonable way to play the game on an appropriate, and greatly simplified, type structure. The converse is also true. Therefore, analysing the game on the right type structure is without loss of generality while being much

Alice	com., t_B^1	com., t_B^2	not, t_B^1	not, t_B^2
t_A^1	0	1	0	0
t_A^2	1/2	0	0	1/2

Bob	com., t_A^1	com., t_B^2	not, t_A^1	not, t_A^2
t_B^1	0	1	0	0
t_B^2	1/2	0	0	1/2

Figure 5. Harsanyi's type structure: For instance, the top table: Alice's types and their beliefs about the genre and Bob's type. For example, when Alice is t_A^1, she assigns probability 1 to Bob being t_B^2 and the feature being a comedy.

simpler. To calculate the equilibrium of the two friends, what is required is to obtain the appropriate choice for each of the four possible types (this is because the type encapsulates a player's view of the world in its entirety).

(f) *Back to the game*

We now settle the question of how the friends would play the game were they to maintain the belief hierarchies generated by figure 2 or, alternatively, figure 5.

Suppose Alice is t_A^2, which would be the case whenever films 1 or 2 are advertised. Recall that t_A^2 assigns probability $\frac{1}{2}$ to the feature being a comedy. Then t_A^2's expected utility from going to the screening is $\frac{1}{2} \cdot 2 + \frac{1}{2} \cdot (-3) < 0$. Because zero is her payoff from staying at home, Alice should stay at home whenever she is t_A^2. On the other hand, she should go to the movie when she is t_A^1 because in that case, she knows the feature is a comedy.

Now suppose Bob is t_B^1. In this case, Bob knows the film is a comedy, but he also knows Alice is t_A^2. By reasoning as we just did, he concludes Alice will stay at home. Clearly, Bob should stay at home whenever he is type t_B^1.

Finally, Bob stays at home when he is t_B^2. To see this note that this type assigns probability $\frac{1}{2}$ to Alice being t_A^1 (she goes to the theatre) and $\frac{1}{2}$ to her being t_A^2 (she stays at home). So t_B^2's expected utility from going is $\frac{1}{2} \cdot 2 + \frac{1}{2} \cdot (-3) < 0$. In sum, Alice goes only to the theatre when 3 is advertised and Bob always stays at home.

(g) *Two questions*

Thus far, we have generated full-fledged theories of mind using compact representations, and used one of these to derive an equilibrium prediction for the incomplete information game.

Two questions remain. The first is: can we go the other way? Suppose, rather than starting from the description in figure 2, we began by considering an arbitrary hierarchy of beliefs. Could we obtain a Harsanyi type structure, for instance, that is consistent with the original hierarchy? After all, type space constructions would be of little use if their employment required extraordinary restrictions on the underlying belief structures. The answer to this first question is yes—provided beliefs satisfy a rather natural consistency requirement [18–20].

The second question concerns the strength of the common knowledge assumption. In order to derive

Alice's hierarchical theory of Bob, it was assumed that figure 2 is common knowledge. One way to justify such common knowledge is to posit that Alice and Bob are familiar with each other's vantage points and that this is really what is common knowledge. Imagine, for instance, that Alice (or Bob) had at some point looked at the marquee from Bob's (or Alice's) vantage point while Bob (or Alice) was there to witness the event. Then common knowledge of the information structure would be natural whenever it was assumed that rational agents come to hold the same beliefs whenever they are fed the same information (this is Harsanyi's [12] basic justification for the assumption). Another way around the problem is to assume that Alice maintains probabilistic beliefs about Bob's information structure. For instance, she might assign probability $\frac{1}{2}$ to his view of the world being described by figure 2 and probability $\frac{1}{2}$ to his having a different vantage point. We could then generate Alice's theory of Bob by assuming this new information structure is common knowledge. Several works have shown that iterating this procedure yields an expanded type space where common knowledge of the knowledge structure can be assumed without loss of generality [20,21].

4. COMPARISON TO OBSERVATIONS AND THEORETICAL NEUROSCIENCE

The earlier-mentioned economic model of *ToM* is an elegant and powerful intellectual achievement, but how does it relate to the theoretical approaches in other disciplines and to empirical data?

What are the neural networks employed by *ToM*? A debate in the neuroscience literature concerns whether *ToM* tasks simply rely on the same networks employed for general logical reasoning, as might seem efficient. Van Overwalle [22] provides a meta-analysis of this question that stresses the distinction in neural activity between *ToM* tasks and general logical reasoning. He finds, in particular, that the medial prefrontal cortex (mPFC) is less often engaged during reasoning tasks that do not involve human agency. Relatedly, Gallagher *et al.* [23] found more activity in the anterior paracingulate cortex (a region previously associated with mentalizing tasks) in subjects that believed they were playing a simple game against a human competitor relative to those subjects that believed they were playing against a computer. Gallagher & Frith [24] review literature implicating the anterior paracingulate cortex, the superior temporal sulci and the temporal poles in *ToM* tasks. In a different setting, Rilling *et al.* [25], found more intense activity in the commonly observed *ToM* neural network when subjects played human, rather than computer, opponents.

Frith & Singer [26] provide a recent review of cognition in social settings. They advocate that specific networks used for processing our own mental states are pressed into service to process the similar mental states of others. Frith [1] argues, more specifically, that mentalizing can occur through an automatic process operating without the protagonist's awareness. That is, we account for the knowledge of others when it is different from our own, but in an effortless, unconscious manner.

There is not a lot of work that explores the fruitful interaction that should be possible between the economic *ToM* and neuroscience. An exception is the work by Yoshida *et al.* [27]. (Yoshida *et al.* [28] investigate the neural mechanisms relevant to [27], and [29] considers the implications for autistic behaviour.) The basic model in these papers considers a repeated game. Each individual makes each choice with a probability that reflects the long-run payoff to that choice. This long-run payoff involves predicting the play of each opponent. A hierarchical structure reminiscent of the *ToM* is obtained by supposing that, to the first order, players ignore the play of their opponents. To the second-order, they suppose their opponents follow a first-order strategy, and so on.

This is very reminiscent of the economic approach of Stahl [30]. Stahl assumes that first-order individuals understand that opponents will not use strategies that are never a best reply. Second-order individuals understand the choices made by first-order individuals, and so on. The key difference between [27] and [30] seems to be the repeated nature of the interaction in the former, and the one-time nature of the interaction in the latter.

People do not merely fail to reason to an infinite depth (order), but do not always adhere to even more basic components of Bayesian rationality. A classic example of such behaviour is due to Ellsberg [31]. A variant of his experiment is as follows. Consider two urns—one urn, R, say with 49 per cent white balls and 51 per cent black balls and another urn, H, say with an unspecified proportion of these two colours. One ball has already been chosen from each urn, but their colours are still unknown by the subject. This subject must nominate either the R-ball or the H-ball. In the first case, the individual wins \$1000 if the ball is black. In the second case, the \$1000 is awarded if the ball is white. Most people choose the R-ball in the first case, suggesting that they estimate that the probability that the H-ball is white is at least 49 per cent. However, most people also prefer the R-ball in the second case as well. There are then no probabilities that can be assigned to the two colours in the ambiguous urn that are consistent with this pattern of choice. This violates 'probabilistic sophistication' or the dictum that choice should depend only on the list of outcomes and the list of associated subjective probabilities for those outcomes.

Hoffrage & Gigerenzer [32] describe experiments demonstrating that medical students and doctors do not estimate probabilities in accordance with Bayes theorem. For example, consider the following question posed to house officers, students and doctors at the Harvard Medical School. A certain disease has an incidence of 1/1000 and there is a test that has a false positive rate of 5 per cent. If the test is positive, what is the probability that the patient actually has the disease? The estimates ranged from 95 per cent (given by 27 out of 60) to the correct answer of 2 per cent (given by 11 out of 60). (This answer is correct assuming a false negative rate of 0%.) This is a little alarming, given that diagnosis is at the heart of the expertise of doctors. The experiment was done in 1978; possibly performance would be better now.

At the same time, there are some startling instances of automatic conformity with Bayes theorem. Ernst & Banks [33] for example, describe how people integrate visual and haptic (touch) sensory inputs about the height of a ridge. If the goal is to minimize the variance of the overall estimate, the visual and haptic inputs should be weighted in inverse proportion to their individual variances. This is a reasonably close description of what actually occurs.

Bayesian rationality serves as a useful benchmark and constitutes an important and fruitful perspective from which to view the data. It would be naive, however, to expect uniformly close agreement between this theory and empirical phenomena. What would be enormously useful as a supplement to Bayesian rationality would be a structured theory of bounded rationality that is not empirically empty.

5. EVOLUTION OF THEORY OF MIND

One promising avenue that is worth investigating in this light is to consider the biological evolutionary genesis of *ToM*. Ultimately, such an approach might suggest a promising way to bound rationality that would not merely fit the data but have out-of-sample predictive power.

Monte *et al.* [34] initiate consideration of this biological genesis of *ToM* from perhaps the most basic point of view possible. It is taken as given that individuals have an appropriate own utility function, so the focus is on the advantage of knowing the utility functions of opponents. Such an advantage is presumed in the other literature in economics. As in the treatment by Robson [8] of the evolutionary advantage of an own utility function, the advantage of knowing another's utility stems from reacting appropriately to novelty. We consider games in which players must learn to play appropriately in an environment that becomes increasingly complex. We show that having a template into which the preferences of an opponent can be fitted enables a sophisticated player to deal with a higher rate of innovation than can a naive individual who adapts to each game.

Consider the argument in greater detail, limiting attention, for the present purpose, to a two-stage extensive form with perfect information. Player 1 moves first, with two choices. Player 2 moves next, again with two choices, but knowing the move made by player 1. In each period, each of a large number of player 1s is randomly matched to an opponent drawn from a large number of player 2s. In addition, the outcomes needed to complete the game are drawn randomly from some large and growing but finite set. Each player has a strict ordering over the set of outcomes. Each player is fully aware of his own ordering but does not know the strict preference ordering of his opponent.

We compare two types of players—naive and sophisticated *ToM* types. In the two-stage setting, this distinction is only important for player 1, because the optimal choice by the player 2s relies only on player 2s' preferences. The naive players behave in a fashion that is consistent with simple adaptive learning in psychology and with evolutionary game theory in economics. Each game is seen as a fresh problem; so naive learners must adaptively learn to play each such different game. For simplicity, however, we assume that this adaptive learning is very fast. The first time a new game arises, the adaptive learner plays inappropriately; but on the second appearance, her play is fully appropriate. This clearly loads the dice against the result we establish concerning the evolutionary advantage to the *ToM* type.

The *ToM* type of player 1, on the other hand, is disposed to learn the other agent's preferences. It is relevant now that the pattern of play is revealed to all players at the end of that period. We assume that each *ToM* type plays inappropriately if she does not know how player 2 will make either pair of choices that might arise in the game. Each time the player 1s see the player 2s being forced to make a choice, the player 1s learn how the player 2s rank the two outcomes. Note here, that the assumption that there are a large number of player 2s means that there is no incentive for the player 2s to choose contrary to the myopic optimum. For simplicity, we do not suppose the player 1s use the transitivity of the preference ordering of the player 2s. Again, this assumption loads the dice against the result we establish concerning the evolutionary advantage of the *ToM* type.

In order to study how relatively successful these two types are, we introduce innovation. The gap between the arrival of new outcomes depends on the existing number of outcomes raised to a power, and it is this power that is the key growth parameter. If the growth rate is low, both types converge on playing appropriately in every game. If the growth rate is high, on the other hand, both types converge on a success rate of zero. For an intermediate range of growth rates, however, the *ToM* types will converge on a success rate of 1, while the naive types will converge on a success rate of zero. In this simple, strong and robust sense, then, the *ToM* type outdoes the naive type. The key reason for the greater success of the *ToM* type is simply that there are vastly more possible games that can be generated from a given number of outcomes than there are outcome pairs.

The two-stage game is special in a number of ways. The player 2s have no need of strategic sophistication at all, and the strategic sophistication of the player 1s is limited to knowledge of the player 2s' preferences. However, we argue that analogous results continue to hold for more general S-stage games of perfect information. The growth in complexity here does not directly stem from higher and higher orders of belief, since when learning about preferences occurs here it is common knowledge. Any growth in complexity that there is stems from the more prosaic need for players moving near the start of the game to obtain preference information about more and more players. Remarkably, this greater complexity does not show up as a decreased ability to respond to novelty.

(a) *A static version of the argument*

We can illuminate the results for the two-stage game by considering a purely static and much simpler version of the argument. Suppose there are N outcomes available

to each player. Consider the naive strategy for player 1 that maps each game to the appropriate action. The complexity of this strategy is then simply the number of such games, which is N^4, because each of the four outcomes of the game has N possibilities.

Consider the *ToM* strategy that anticipates the choice made by player 2 between each possible pair of outcomes for player 2. The complexity of this strategy is the number of such pairs, namely $N(N-1)/2$, which is only second-order in N.

It is plausible that the *ToM* strategy entails some additional computational cost, associated with an additional computational procedure. This additional cost might then well be independent of N, however. Even with a fixed cost component, the *ToM* strategy will be preferred, at least for large enough N.

(b) *Testing the model*

It would be of inherent interest to experimentally implement a version of the model outlined earlier, perhaps simplified to have no innovation: that is, put a reasonably large number of subjects into each of two pools—one for the player 1s and one for the player 2s; induce the same preferences over a large set of outcomes for each of the player 1s and for each of player 2s by using monetary payoffs, but where neither side knows the other side's preferences; and play the game otherwise as outlined already.

How fast would the player 1s learn the player 2s' preferences? Would they be closer to the sophisticated *ToM* types described earlier or to the naive types? How would children play? How would autistic individuals play? What regions of the brain would be activated in solving this task? (Some of the likely candidates for neural networks to be activated are discussed in §4.) Would individuals who are more sophisticated, in the sense of using the information available more efficiently, show greater activation in some specific regions?

6. CONCLUSION

This collection is devoted to perhaps the most startling of all evolutionary products—human cognition. We focus on an aspect that seems particularly likely to contribute to understanding our prodigious sociality and resultant evolutionary success—*ToM*. Although the economic theory presented here is powerful, elegant and general, it does not closely fit the data. There is the promise here of fruitful interdisciplinary interaction to develop a useful theory of bounded rationality. One promising avenue of investigation is to consider the biological basis of decision theory and game theory. We accordingly outline two papers that consider the biological basis of *ToM*, as a start to developing a less sophisticated, more realistic, but still general, theory.

Robalino's research was funded by the Human Evolutionary Studies Program at Simon Fraser University. Robson's research was funded by a Canada Research Chair and by the Human Evolutionary Studies Program at Simon Fraser University. We thank the participants in various conferences and seminars and the referees for useful comments and suggestions.

REFERENCES

1 Frith, C. D. 2012 The role of metacognition in human social interactions. *Phil. Trans. R. Soc. B* **367**, 2213–2223. (doi:10.1098/rstb.2012.0123)
2 Klemperer, P. 2004 *Auctions: theory and practice*. Princeton, NJ: Princeton University Press.
3 Samuelson, P. 1938 A note on the pure theory of consumer behavior. *Economica* **5**, 61–71. (doi:10.2307/2548836)
4 Mas-Colell, A., Whinston, M. D. & Green, J. R. 1995 *Microeconomic theory*. Oxford, UK: Oxford University Press.
5 Binmore, K. G. 1990 *Essays on the foundations of game theory*. Oxford, UK: Blackwell.
6 von Neumann, J. & Morgenstern, O. 1944 *Theory of games and economic behavior*. Princeton, NJ: Princeton University Press.
7 Savage, L. J. 1954 *The foundations of statistics*. London, UK: John Wiley and Sons.
8 Robson, A. J. 2001 Why would Nature give individuals utility functions? *J. Political Econ.* **109**, 900–914. (doi:10.1086/322083)
9 Nash, J. F. 1950 Equilibrium points in n-person games. *Proc. Natl Acad. Sci. USA* **36**, 48–49. (doi:10.1073/pnas.36.1.48)
10 Nash, J. F. 1951 Non-cooperative games. *Ann. Math.* **54**, 286–295. (doi:10.2307/1969529)
11 Harsanyi, J. 1962 Bargaining in ignorance of the opponent's utility function. *J. Conflict Resolution* **6**, 29–38. (doi:10.1177/002200276200600104)
12 Harsanyi, J. 1967–1968 Games with incomplete information played by 'Bayesian' players, I–III. *Manag. Sci.* **14**, 159–182, 320–334, 486–502. (doi:10.1287/mnsc.14.3.159)
13 Aumann, R. J. & Heifetz, A. 2002 Incomplete information. In *Handbook of game theory with economic applications*, vol. 3 (eds R. J. Aumann & S. Hart), pp. 1665–1686. Amsterdam, The Netherlands: North-Holland.
14 Aumann, R. J. 1987 Correlated equilibrium as an expression of Bayesian rationality. *Econometrica* **55**, 1–18. (doi:10.2307/1911154)
15 Aumann, R. J. 1976 Agreeing to disagree. *Ann. Stat.* **4**, 1236–1239. (doi:10.1214/aos/1176343654)
16 Geanakoplos, J. 1992 Common knowledge. *J. Econ. Perspect.* **6**, 53–82.
17 Samuelson, L. 2004 Modeling knowledge in economic analysis. *J. Econ. Lit.* **42**, 367–403. (doi:10.1257/0022051041409057)
18 Armbruster, W. & Boge, W. 1979 Bayesian game theory. In *Games and related topics* (eds O. Moeschlin & D. Pallaschke), pp. 17–28. Amsterdam, The Netherlands: North-Holland.
19 Boge, W. & Eisele, Th. 1979 On solutions of Bayesian games. *Int. J. Game Theory* **8**, 193–215. (doi:10.1007/BF01766706)
20 Mertens, J.-F. & Zamir, S. 1985 Formulation of Bayesian analysis for games with incomplete information. *Int. J. Game Theory* **14**, 1–29. (doi:10.1007/BF01770224)
21 Brandenburger, A. & Dekel, E. 1993 Hierarchies of belief and common knowledge. *Games Econ. Behav.* **59**, 189–198.
22 Van Overwalle, F., 2011 A dissociation between social mentalizing and general reasoning. *NeuroImage* **54**, 1589–1599. (doi:10.1016/j.neuroimage.2010.09.043)
23 Gallagher, H. L., Jack, A. I., Roepstorff, A. & Frith, C. D. 2002 Imaging the intentional stance in a competative game. *NeuroImage* **16**, 814–821. (doi:10.1006/nimg.2002.1117)
24 Gallagher, H. L. & Frith, C. D. 2003 Functional imaging of 'theory of mind'. *Trends Cogn. Sci.* **7**, 77–83. (doi:10.1016/S1364-6613(02)00025-6)

25 Rilling, J. K., Sanfey, A. G., Aronson, J. A., Nystrom, L. E. & Cohen, J. D. 2004 The neural correlates of theory of mind within interpersonal interactions. *NeuroImage* **22**, 1694–1703. (doi:10.1016/j.neuroimage.2004.04.015)
26 Frith, C. D. & Singer, T. 2008 The role of social cognition in decision making. *Phil. Trans. R. Soc. B* **363**, 3875–3886. (doi:10.1098/rstb.2008.0156)
27 Yoshida, W., Dolan, R. J. & Friston, J. K.. 2008 Game theory of mind. *PLoS Comput. Biol.* **4**, e1000254. (doi:10.1371/journal.pcbi.1000254)
28 Yoshida, W., Seymour, B., Friston, K. J. & Dolan, J. R. 2010 Neural mechanisms of belief inference during cooperative games. *J. Neurosci.* **30**, 10 744–10 751. (doi:10.1523/JNEUROSCI.5895-09.2010)
29 Yoshida, W., Dziobek, I., Kliemann, D., Heekeren, H. R., Friston, K. J. & Dolan, R. J. 2010 Cooperation and heterogeneity of the autistic mind. *J. Neurosci.* **30**, 8815–8818. (doi:10.1523/JNEUROSCI.0400-10.2010)
30 Stahl, O. D. 1993 Evolution of smart$_n$ players. *Games Econ. Behav.* **5**, 604–617. (doi:10.1006/game.1993.1033)
31 Ellsberg, D. 1961 Risk, ambiguity, and the savage axioms. *Q. J. Econ.* **75**, 643–669. (doi:10.2307/1884324)
32 Hoffrage, U. & Gigerenzer, G. 2004 How to improve the diagnostic inferences of medical experts. In *Experts in science and society* (eds E. Kurz-Milcke & G. Gigerenzer), pp. 249–268. New York, NY: Kluwer Academic/Plenum Publisher.
33 Ernst, M. O. & Banks, M. S. 2002 Humans integrate visual and haptic information in a statistically optimal fashion. *Nature* **415**, 429–433. (doi:10.1038/415429a)
34 Monte, D., Robalino, N. & Robson, A. J. 2012 The evolution of 'theory of mind'. Working Paper; Economics Department, Simon Fraser University, Burnaby, Canada.