Introduction

Samir Okasha and Ken Binmore

There exist deep and interesting connections, both thematic and formal, between evolutionary theory and the theory of rational choice, despite their apparently different subject matters. These connections arise because a notion of optimization or maximization is central to both areas. In rational choice theory, agents are assumed to make choices that maximize their utility, while in evolutionary theory, natural selection 'chooses' between alternative phenotypes, or genes, according to the criterion of fitness maximization. As a result, evolved organisms often exhibit behavioural choices that appear designed to maximize their fitness, which suggests that the principles of rational choice might be applicable to them. This conceptual link between evolution and rational choice explains the fascinating exchange of ideas between evolutionary biology and economics that has taken place in the last forty years, particularly in relation to decision making under uncertainty, and strategic interaction.

The chapters in this book all deal with aspects of the evolution/rationality relationship, from a range of perspectives. The book emerged from a series of workshops and conferences held at the University of Bristol between 2008 and 2011, under the auspices of the 'Evolution, Cooperation and Rationality' research project, funded by the Arts and Humanities Research Council of the UK and directed by ourselves. The project examined foundational and conceptual issues arising from recent work on social behaviour, decision making and strategic interaction, and had a strongly interdisciplinary orientation. This is reflected in the composition of the book – the authors include leading researchers in evolutionary biology, philosophy of science, experimental economics, game theory and psychology. The result illustrates the rich diversity of approaches to the study of evolution and rationality, and, we hope, will help promote constructive dialogue between them.

The fundamental paradigm of the economic theory of rational choice is that someone who chooses 'consistently', where this means conforming to

certain rather intuitive axioms, behaves as though maximizing expected utility. Biologists similarly argue that in suitably idealized circumstances, evolution will produce animals that behave as though maximizing their expected fitness, where 'fitness' refers to the additional number of offspring the animal produces as a result of the behaviour in question. So the question arises: when it is possible to identify the economists' notion of utility with the biologists' notion of fitness?

This question of the relationship between utility and fitness is a central theme in a number of chapters here, and is in the background of most of the others. Kim Sterelny defends the idea that fitness maximization and utility maximization can sometimes be expected to coincide in human populations, but not always. It depends on whether information is transmitted vertically or horizontally, and on whether group selection is or is not a prevalent factor, Sterelny argues, for these determine whether population-level processes will be effective in binding proximal motivation to fitness consequences. The utility versus fitness issue is also central to the chapter by Claire El Mouden, Maxwell Burton-Chellew, Andy Gardner and Stuart A. West, who ask what quantity – if any – we should expect humans to appear designed to maximize. Their analysis is based on inclusive fitness theory, the highly successful approach to social behaviour devised by W. D. Hamilton. El Mouden et al. make the case that humans, like other social animals, have been selected to maximize their inclusive fitness. They admit that much human behaviour doesn't appear to actually achieve this, but consider a number of reasons, quite different from Sterelny's, for how to square the data with the assumption of fitness maximization.

Alasdair Houston's chapter also tackles the issue of fitness and utility, from a very different perspective. A key assumption in rational choice theory is that an agent's preferences or choices should be transitive; otherwise, the agent cannot be represented as a utility maximizer. However, systematic intransitivities of choice have been reported in both human and non-human subjects; this raises the question of how such apparently irrational behaviour could have evolved. Houston examines a number of potential explanations for how natural selection can result in intransitive choices. Some of these explanations, though not all of them, imply that the choice behaviour is only apparently irrational – in that transitivity can be restored so long as the modeller takes a 'correct view' of the decision maker's options. In particular, Houston stresses the importance of considering 'state-dependent' decisions, in which an animal's choice behaviour is partly determined by a state variable, e.g. its energy reserves.

If we neglect state dependence, we may be led to see an animal's choice behaviour in any one state as irrational, when in fact it is implementing an optimal state-dependent strategy.

The idea that there is a fairly straightforward evolutionary basis for rational choice maxims such as consistency of preferences and maximization of expected utility is developed in this volume by Herbert Gintis, in his chapter on the unification of the behavioural sciences from an evolutionary perspective. Gintis defends the 'rational actor' model that underpins most economic theory, and argues that it integrates seamlessly with evolutionary biology. He rejects the view, held by many psychologists, that humans exhibit systemic cognitive biases and irrationalities which undermine the applicability of rational choice theory. General evolutionary considerations tell against this pessimistic idea, Gintis argues, and the experimental data can be explained in ways that are compatible with the rational actor model. However, a complete theory of behavioural choice must go beyond the rational actor model, he thinks, and incorporate ideas from both evolutionary biology and social psychology. Gintis outlines how he thinks this conceptual unification should take place.

A quite different attitude towards rational choice theory is found in Henry Brighton and Gerd Gigerenzer's chapter, who focus on a problem faced by all organisms: making inductive inferences in an uncertain world. Their chapter builds on Gigerenzer's previous work in which he strongly criticizes the use of optimality and rational choice models to understand adaptive behaviour, arguing that 'simple heuristics' will often outperform attempts at maximization. Brighton and Gigerenzer defend this view in relation to inductive inference. They argue that rational choice models only work in what L. J. Savage called 'small worlds', i.e. situations where the state space of the decision problem is pregiven, but are highly misleading in 'large worlds', where the state space must itself be inferred from observations. In large worlds, adaptive behaviour is easier to produce via heuristics than optimization. Gigerenzer and Brighton connect this argument with an interesting philosophical claim, namely that there is no such thing as a 'one true rationality', since rationality principles are invented, not discovered. They suggest an understanding of 'rational' and 'optimal' which is compatible with this philosophy.

An area where the interplay of ideas between economics and evolution has been particularly fruitful is game theory. Originally designed to explain the strategic choices of rational human agents, game theory was introduced into biology in the 1970s and 1980s to explain aspects of animal behaviour. The basic concept of traditional game theory is the idea

of a Nash equilibrium. A profile of strategies – one for each player – is a Nash equilibrium if no player has an incentive to deviate from his strategy provided that nobody else deviates first. In traditional game theory, the Nash equilibrium is interpreted as an equilibrium in rational deliberation, i.e. a situation from which no rational player will unilaterally deviate. Thus we can predict that rational players in a game will end up at a Nash equilibrium of that game. However, the Nash equilibrium also admits of an evolutionary interpretation, because any dynamical process that always moves in the direction of higher payoffs can only stop when it gets to a Nash equilibrium.

This dual interpretation has proved very useful in evolutionary biology, because it sometimes allows theorists to use the rational interpretation to predict the outcome of an evolutionary process without needing to study the complicated details of the process itself. When reasoning in this way, biologists usually speak of an *evolutionarily stable strategy* (ESS), a concept first devised by John Maynard Smith and George Price and which bears a close relation to the Nash equilibrium concept. An ESS refers to a state of a population that, once reached, cannot be invaded by small groups of mutants.

Peter Hammerstein's chapter traces the fascinating intellectual history of how game theory entered evolutionary biology, with a focus on conceptual and foundational issues. He makes a strong case for the power of strategic analysis in biology, citing numerous examples of biological phenomena that have been illuminated through the application of game-theoretic methods, including conflicts over parental investment and intraorganismic conflict. Despite these success stories, and despite the fact that biological game theory can dispense with the improbably strong epistemic assumptions made by classical game theory – such as common knowledge of rationality – Hammerstein sounds a note of caution. Biologists cannot simple appeal to the 'authority of traditional game theory' to analyse the consequences of strategic interaction, he argues; explicit attention to the evolutionary dynamics, rather than merely looking for stable equilibria, may be required.

This issue of evolutionary dynamics, and the relation between rational and evolutionary game theory, are also central to the chapter by Simon M. Huttegger and Kevin J. S. Zollman. They offer a searching critique of what they call 'ESS methodology' in biology, i.e. the practice of assuming that evolution will take a population to an ESS, and using this to guide the interpretation of observed biological phenomena. The problem with this methodology is that in some circumstances, natural selection

will not carry a population to an ESS state, so the methodology is at best fallible; there is no short cut to studying the full evolutionary dynamics, they argue. Huttegger and Zollman offer a striking illustration of this point with a simple 'sender–receiver' signalling game, in which most initial population states converge to an equilibrium that is not an ESS. Interestingly, Huttegger and Zollman trace the failure of the ESS methodology to a feature that the ESS concept shares with other 'refinements' (or logical strengthenings) of the Nash equilibrium concept discussed in traditional rational game theory.

Many authors have contrasted rational and evolutionary game theory, or viewed them as rival 'interpretations' of an underlying formalism. An alternative approach is to try to integrate the two. This is the route taken by Siegfried Berninghaus, Werner Güth and Hartmut Kliemt, who advocate what they call an 'indirect evolutionary approach'. (Somewhat similar ideas have also been developed under the heading 'evolution of preferences'.) The core of Berninghaus et al.'s indirect evolutionary approach is to model strategic behaviour using two timescales, long and short, corresponding to ultimate and proximate levels of explanation. In the short term, agent's subjective preferences, and thus their behavioural choices, may be governed by rules and social norms, rather than by the pursuit of Darwinian fitness. But over a longer evolutionary timescale, the proliferation of different behaviours depends on their 'objective payoffs', or fitnesses. Berninghaus et al. show that subjective motives other than maximization of objective payoff can be favoured (a result that tallies with Sterelny's argument). The indirect evolutionary approach offers an interesting take on the relationship between agents' subjective motivations and the evolutionary consequences of their actions.

The 'two-timescales' idea also features in the chapter by David H. Wolpert and Julian Jamison, on the strategic choices of 'non-rational' players. In their version, however, the longer timescale corresponds to learning within the lifetime of a single player, rather than an evolutionary process unfolding over multiple generations. (It is well known that evolution and learning exhibit interesting parallels, a point discussed by Hammerstein.) Focusing on learning rather than evolution permits Wolpert and Jamison to stick with the Nash equilibrium concept, rather than the ESS concept which is harder to work with. Wolpert and Jamison's central idea is that of a 'persona game', in which a player chooses a persona, e.g. that of someone who refuses to be treated unfairly, signals the persona to others, and commits to using it during the play of a game. (This need not be done fully consciously.) Wolpert and Jamison argue that humans have a

remarkable ability to adopt different personae in their social interactions, and explore the subtle strategic implications of this ability. They show how various forms of 'non-rational' behaviour, such as co-operation in a one-shot prisoner's dilemma, can be explained via persona games, and argue that this explanation fits the extant data.

Co-operative behaviour, and the problem of how to incorporate it into a systematic framework, is also central in Natalie Gold's chapter. She explores the notion of 'team reasoning', originally due to Michael Bacharach and Bob Sugden, as a potential explanation of apparently non-rational choices in games such as the prisoner's dilemma. In standard game theory, the players reason in an 'individualistic' way, aiming to maximize their own utility (though of course their utility function may be 'other regarding'). In team reasoning, players are able to mentally identify with a particular team, or set of players, and make choices that are optimal from the point of view of the whole team. Gold explores two subtly different versions of team reasoning, and shows how it can lead players to co-operate in social dilemmas. It is particularly interesting that the basic idea behind team reasoning – invoking 'team payoff' in addition to individual payoff – can also be found in evolutionary biology, in the theory of multilevel selection, a point that Gold discusses.

Co-operation and social behaviour are also central to the chapter by Jack Vromen, which is a philosophical investigation of recent work on 'strong reciprocity' in humans. Strong reciprocity refers to our predisposition to co-operate with others, and to punish others who fail to co-operate, even when this punishment is costly to administer. There is considerable evidence, both experimental and anthropological, in favour of the idea that humans are strong reciprocators in this sense. Vromen argues that the literature on strong reciprocity contains a number of conceptual confusions, in particular over whether it constitutes altruism or selfishness, and whether it requires group-level selection to evolve. He traces these confusions to a failure to keep separate the evolutionary and the psychological meanings of 'altruism', an issue closely connected to the distinction between proximate and ultimate explanation. He also examines the psychological evidence on our propensity to engage in costly punishment, arguing that it cannot resolve the issue of whether this propensity is psychologically selfish or altruistic.

This work was supported by the Arts and Humanities Research Council, grant no. AH/FO17502/1, and the European Research Council Seventh Framework Program (FP7/2007–2013), ERC Grant agreement no. 295449.

CHAPTER 1

Towards a Darwinian theory of decision making
Games and the biological roots of behavior

Peter Hammerstein

1.1 INTRODUCTION

Conventional decision theory is normative and it attempts to identify decisions that are in some sense optimal. The decision maker is often assumed to have all the mental capabilities that real human beings can only dream of. Classical game theory has built on this approach and many of its scholars have almost routinely referred to the normative character of their theory as an excuse for the lack of empirical content. I claim that this excuse is unconvincing. Even an entirely rational visitor from outer space would meet real people on earth and would have to deal with them in a smart way. This visitor would be forced to learn as much as possible about the evolved psychology of humans in order to identify his best decisions in our world of the not-so-smart. It is therefore impossible to separate normative and descriptive approaches unless game theory deals exclusively with rational visitors from outer space.

In this chapter, I wish to explain how game theory can be firmly rooted in the life sciences without dismissing the legacy of its founders. I first take a look at the history of ideas in game theory and give my comments as a theoretical biologist. The next step is to explain the interesting links between reasoning in decision theory and those properties of the evolutionary process that look to us *as if* evolution itself were able to reason about decision problems. A subsequent excursion into the bacterial world demonstrates that even microbes reflect this feature of evolution. Looking finally at animal interactions, I discuss how basic ideas in game theory sometimes hit the nail on the head in relation to empirical findings and are at other times very misleading. A concluding remark is devoted to learning and the future of game theory.

I am grateful to Benjamin Bossan and Arnulf Koehncke for numerous comments I received when preparing this chapter.

1.2 A BIOLOGIST'S LOOK AT THE STRUGGLE FOR CONCEPTS IN CLASSICAL GAME THEORY

As a scientific discipline game theory emerged early in the twentieth century and gained general visibility in 1944 when John von Neumann and Oskar Morgenstern published their seminal book *Theory of Games and Economic Behavior*. The new discipline was meant to provide a framework for mathematical modeling in economics and the social sciences. Its development was, therefore, driven by the need to capture the essentials of decision making in interactive situations. Obviously, this need could not be satisfied by simply borrowing ideas from physics or any of the other natural sciences. Conversely, game theory later became the first formalized field of the social sciences that had considerable impact on theory development in a natural science. This was the case when evolutionary game theory emerged in biology (Maynard Smith and Price 1973; Maynard Smith 1982; Hammerstein and Selten 1994).

A game is, technically speaking, a mathematical model of an interaction with two or more actors (players) involved. Von Neumann and Morgenstern introduced basic forms of models that are still in use, such as the extensive and normal (strategic) form of a game. The founders of game theory were less successful, though, in foreshadowing the solution concepts (ways of analyzing games) that later became mainstream practice. Why did they not anticipate more of the theory for which they had paved the road? Undoubtedly, John von Neumann was a person with great visionary power and exactly for this reason he probably appreciated the difficulty of the following fundamental question: What can be regarded as a player's adequate strategic response to the strategy choices of other players if – as is usual – those choices are *unknown*? (Note here that a strategy is generally far more than a simple observable act, and even the choice of a simple act can only be observed after it has taken place.)

In my view, it is this *unknown* that should have puzzled other game theoreticians more than it typically did. Von Neumann demonstrated his appreciation of the strategic response problem by introducing a solution concept that avoids speculating about the unknown. Unfortunately, this avoidance led him to a very pessimistic decision principle, which is to "minimize maximum losses" by playing so-called *minimax strategies*. His minimax solution concept seems unconvincing from a modern point of view except for the context of two-person zero-sum games. The main criticism is that "minimaxers" ignore how likely it is that the persons they

interact with will choose different behavioral options. Here, a defender of minimax could reply that *uncertainty* exists with regard to these probabilities. The defender would perhaps claim that one should take into account only those probabilities that are known, like that of rolling a given number with a single throw of dice. Only in this case of known probabilities is the risk calculable. This attempt to steer the conceptual discourse is debatable, however, since it assumes a clear distinction between decision under risk and uncertainty. At least from a Bayesian point of view such a distinction cannot be made, and rationality axioms in the footsteps of Savage (1954) would force a decision maker to form subjective probabilities in every decision situation.

The beauty of Bayesian decision theory should not prevent us from realizing that in most real-life situations it seems technically impossible for humans to practice with their evolved brains what Bayesian axioms would require them to do. Selten (1991) therefore expressed the view that Bayesian theory can neither describe typical human decision making, nor can it frequently be of practical normative use. He only admits that the application of Bayesian methods can make sense in special contexts. In the information processing of an insurance company, for example, subjective probabilities may be generated reasonably well on the basis of actuarial tables.

As a scientist trying to capture reality I can hardly disagree with Selten's radical view on Bayesian concepts but also have to admit – as does Selten – that one can, if one wishes, think about fictitious rational beings that possess all the technical expertise and capacities needed for Bayesian decision making. As a thought experiment it may then be feasible to explore the interactions that could take place in this fictional world. But caution is more than strongly advised when returning to the world of facts.

Addressing now the most successful solution concept of classical game theory, the Nash equilibrium (Nash 1951), we run into the same problems as before. A Nash equilibrium specifies strategies for each player in such a way that no player could improve his expected payoff by unilaterally deviating from this specified strategy profile. Nash's concept looks rather trivial at first glance, when one realizes that it merely transfers the idea of optimization (here maximization of payoff) to interactive situations (every player responds optimally to the other players), in a simple and intuitive way. This may well be the reason why Nash as a mathematician considered it an obvious choice. His concept raises many questions, however, if one starts thinking about its deeper justification.

What assumptions are needed to back up the Nash equilibrium? This depends strongly on the perspective taken. Let us start by asking how a player could find a Nash equilibrium through some kind of careful reasoning rather than by intuition, routine learning, or teaching. In order to anticipate the actions of others, a carefully reasoning player would have to theorize about their minds, and – unfortunately for empirically void, idealized approaches – the minds of others may not be operating through careful reasoning. A sensible way of conducting one's thought would, therefore, be to rely on empirical knowledge about the evolved psychology of decision making. From behavioral experiments we know that this evolved psychology often does not favor Nash equilibria. When interacting with real people, rational players would thus frequently be forced to avoid Nash equilibria in order to play best responses to the strategies that actually matter, and that they actually encounter. Looked at from this angle, the Nash equilibrium fails to be convincing even as a normative concept: *we often ought not to do what classical game theory claims we ought to do*.

Now, in order to come up with a fairly general, reasoning-based justification of the Nash equilibrium, one has to invoke something like the following grandiose assumption:

(A1) All players in a game are rational and all know that all know that all are rational.

Why does this assumption help? Since all players are now artificial beings, void of psychology, and slaves of the axioms of rationality, they have no problem figuring out how everybody else will make strategic choices. "Reading the mind" can here simply be replaced by "reading the axioms." Consequently, if these rational players adopt a solution of how to play the game, this solution must be consistent with the axioms that define their fictitious minds. No solution can then be adopted that includes payoff incentives for deviation. Along this line of reasoning we can interpret the Nash equilibrium property as a *necessary condition for what qualifies as a rational decision in a purely rational world*.

Note that in special cases the Nash equilibrium is justified under assumptions weaker than (A1). For example, in games with strictly dominant strategies, such as the prisoner's dilemma, an optimal strategic decision can be made without assuming anything about the rationality or the knowledge of others. Note also that a group of players educated in classical game theory, all knowing about their joint education and trusting the success of "brainwashing" by their teachers, may