SAMIR OKASHA and KEN BINMORE (eds) Evolution and Rationality: Decisions, Cooperation, and Strategic Behaviour. Cambridge: Cambridge University Press, 2012, 296 pp., £60 (hardback) ISBN: 978-1107004993

Jonathan Birch

Evolution and Rationality marks the end of a three-year project, 'Evolution, Cooperation, and Rationality', directed at the University of Bristol by the book's editors, Samir Okasha and Ken Binmore. The collection draws together the editors' pick of the papers delivered at the conferences the project hosted, and covers a wide range of topics at the intersection of evolutionary theory and the social sciences. It is a splendid anthology: timely, interdisciplinary, thematically cohesive, and full of substantive and interesting disagreements between the contributors.

The motivation behind the book—and its corresponding project—is the observation that both economics and evolutionary biology make heavy use of what might be termed a 'maximizing-agent model' of behaviour. Economists often model humans as rational maximizers of a quantity called utility, while biologists (particularly behavioural ecologists) often model organisms as if they were maximizers of a quantity called inclusive fitness. Given this parallel, there is obvious scope for two-way discussion regarding the uses and limits of such models, and regarding the meaning and measurement of the quantities that are taken to be maximized. I cannot do justice here to all the specific topics that these articles address, nor can I weigh into all the disputes I would like to weigh into. Instead, I will simply highlight two of the most important issues that this collection raises. One is the fate of the rational actor model of human behaviour. The other is the explanatory role of inclusive fitness in the biological and social sciences.

1 The Rational Actor Model

The question that most strongly divides the contributors is that of the continuing utility (or otherwise) of the rational actor model of human behaviour embodied in contemporary rational choice theory. The case against is supplied most forcefully by Henry Brighton and Gerd Gigerenzer, whose argument draws on a distinction (first proposed by Savage [1954]) between 'large' and 'small' worlds. This distinction has nothing directly to do with the size of social groups; it rather concerns the complexity of the decision problems agents face, and their epistemic access to relevant information. The idea, in broad terms, is that an agent in a 'small world' faces a limited number of strategic options and has some knowledge of the likely consequences of each action, whereas an agent in a 'large world' has a space of options too big to fully comprehend and is largely ignorant of the consequences any particular decision will have. Savage claimed that Bayesian decision theory provides a substantive theory of rational decision making in small worlds, but does not scale up to large worlds: with limitless possibilities and uncertain consequences, the rational choice is left radically indeterminate. Brighton and Gigerenzer update Savage's arguments using a barrage of methods from modern theories of statistical learning. They proceed to argue for a 'relativistic approach' to rational decision making in large worlds, on which the appropriate application of context-specific heuristics is taken to be constitutive of rationality.

No one provides an unqualified defence of the traditional rational actor model in light of such critiques; but several contributors do argue that, provided its limitations are adequately taken into account, the model still has its uses. The contributions by Herbert Gintis and Kim Sterelny are the best examples of this. Sterelny too distinguishes between small and large worlds, although here the key distinction concerns the size of social networks rather than the complexity of decision problems. In contrast to Brighton and Gigerenzer, Sterelny argues that larger worlds make it more appropriate to model humans as rational economic agents, roughly because in such worlds, opportunities for reciprocity or kin-directed preference become comparatively less common, and one-off transactions with strangers become comparatively more so. In the transition from the Pleistocene to the Holocene, Sterelny argues, we see important changes in both the motivational architecture and information-sharing habits of human agents, eventually 'reshaping the extent to which individual decision making tracks fitness' (p. 262). It is not entirely clear to me why Sterelny talks of a transition 'from fitness to utility' in this context, as the notion of utility-maximization seems equally applicable at both ends of the transition, given a suitably broad notion of utility. It seems rather that we have a transition from a Pleistocene social world in which fitness and utility are mostly well aligned to a Holocene social world in which they are often radically misaligned.

Gintis, meanwhile, provides a whistle-stop tour through his 'unification of the behavioural sciences' programme, outlined in greater depth and detail in his ([2009]) book. Although he is ultimately supportive of the rational actor model, his support is not unequivocal. The model, he suggests, is more or less on the right lines and can be underwritten by evolutionary considerations, provided fitness and utility do not come apart too dramatically. But we know from evolutionary theory that good fitness maximizers will sometimes act in ways that are at odds with their economic or psychological self-interest, narrowly construed. So the viability of the rational actor model depends on our ability to pin down a notion of utility that is broader than this, but that still retains enough empirical content to be of explanatory value.

2 What Exactly Is Maximized in Evolution?

I turn now to a different cluster of issues, situated closer to the philosophy of biology than to the philosophy of social science. Debates about the nature of fitness are, of course, time-honoured journal fodder in the philosophy of biology. It is striking, however, that the notion of inclusive fitness has received relatively little philosophical attention in its own right, even though it is inclusive fitness—not individual fitness—that organisms typically appear as if maximizing in social contexts (Hamilton [1964]; Grafen [2006]).

Perhaps we owe this comparative neglect to the fact that inclusive fitness is often thought to be a simple extension of individual fitness: a case of taking an organism's fitness and augmenting it with the fitness of its collateral relatives, weighted by a fraction that quantifies the closeness of their kinship (0.5 for full siblings, 0.25 for nieces and nephews, 0.125 for cousins, and so on). On closer inspection, however, the notion of inclusive fitness turns out to be anything but simple, and if we computed it in the way outlined in the previous sentence, we would quickly run into serious errors (Grafen [1982]). From a philosophical point of view, the most interesting feature of inclusive fitness is its inherently causal character: when defined correctly, it is a weighted sum of the fitness effects for which a particular social actor is causally responsible. One consequence of this is that only relatives with whom an actor has causally interacted can make a difference to its inclusive fitness. In other words, having a successful relative will enhance one's inclusive fitness only if one is causally responsible for a portion of that relative's success (Grafen [1982]). Another more subtle consequence is that an inclusive fitness calculation is not simply a matter of augmenting an individual's fitness with additional components from collateral relatives; we must also subtract from an individual's fitness all those components for which another agent was causally responsible (Hamilton [1964]). The alternative would be a dangerous form of double counting: each social fitness effect would be counted twice, once in calculating the inclusive fitness of the actor and again in calculating that of the recipient.

I raise these subtleties here because they are at times neglected in this volume, and in the evolutionary literature more generally. Take, for instance, Maynard Smith's ([1991]) 'Sir Philip Sidney' game, analysed here by Simon Huttegger and Kevin Zollman. This is a signalling game in which a donor either offers or declines to offer water to a beneficiary who is either thirsty or not thirsty, and who can choose among three signalling strategies: always signal, never signal, or signal only when thirsty. If we want to see informative signalling, it is essential that the payoffs are computed in inclusive fitness terms—otherwise, the donor would never offer the water and the beneficiary would never signal its thirst. The trouble is that the payoffs in this game seem to exhibit precisely the kind of double counting that Hamilton warned against: each agent's survival probability is counted twice, once towards the inclusive fitness of the donor and again towards that of the beneficiary. This is not to say that the game does not shed any light on the evolution of signalling; but if it does, this can only be because the miscalculations of inclusive fitness it embodies turn out to be harmless.

The three-way relationship between inclusive fitness theory, game theory, and decision theory remains ripe for philosophical investigation. In particular, the scope and limits of inclusive fitness theory as a tool for explaining human behaviour remain poorly understood. In light of this, perhaps the most frustrating paper in the volume is Claire El Mouden and colleagues' contribution, 'What do Humans Maximize?'. The question arises in the first place because humans are unusual animals, in that our behaviour has not just been shaped by natural selection; it is at least as much the product of cultural evolution, and of gene-culture co-evolution (Richerson and Boyd [2005]; Gintis [2009]; Sterelny [2012]). This naturally leads us to ask whether we can find some appropriate quantity—some expanded conception of inclusive fitness, perhaps—that can still be said to be maximized overall when genetic and cultural evolution take place concurrently.

This is an important question, and an open one. Unfortunately, El Mouden et al. do not seriously engage with it. Instead, they insist that the 'ultimate goal' of human behaviour really is the maximization of inclusive fitness in the traditional sense:

How would an alien biologist that was capable of observing our behaviours and reading our minds sum up what humans maximize? […] While finding them fascinating for many reasons, the alien would conclude that humans, along with all other organisms, are best described as striving to maximize their inclusive fitness over their lifetimes, yet in imperfect and non-optimal, but often predictable, ways. (p. 42)

It is tempting here to start listing counterexamples: instances in which inclusive fitness maximization is not merely 'imperfect and non-optimal', but radically at odds with an agent's manifest behaviour. Yet, El Mouden et al. have a one-size-fits-all response to such cases. They concede that culture makes things messy, and that it often prevents us from attaining inclusive fitness optima, and that our culturally influenced conscious intentions don't always line up with our evolutionary design objective. But they maintain that such cultural influences are merely proximate causes and that, to reveal the ultimate goal or purpose of human behaviour, one is entitled to abstract away from them. The reader is left feeling that an opportunity for critical reflection on the explanatory limitations of inclusive fitness theory has been missed here; and worse, that Mayr's proximate–ultimate distinction is being misappropriated to shut down an important debate before it has really begun.

This is by no means the only chapter that will polarize readers. On the contrary, this volume is packed with forthright statements of controversial positions. Yet in compiling such varied (and often conflicting) visions of human behaviour in a single volume, the editors have produced a book that, taken as a whole, comes across as pluralistic and open-minded. The overarching message of Evolution and Rationality—to the extent that there is one—is that the task of explaining human behaviour calls for the integration of highly diverse approaches drawn from every corner of the social and biological sciences.

References

Gintis, H. [2009]: The Bounds of Reason: Game Theory and the Unification of the Behavioral Sciences, Princeton, NJ: Princeton University Press.

Grafen, A. [1982]: 'How Not to Measure Inclusive Fitness', Nature, 298, pp. 425–6.

Grafen, A. [2006]: 'Optimization of Inclusive Fitness', Journal of Theoretical Biology, 238, pp. 541–63.

Hamilton, W. D. [1964]: 'The Genetical Evolution of Social Behaviour', Journal of Theoretical Biology, 7, pp. 1–52.

Maynard Smith, J. [1991]: 'Honest Signalling: The Philip Sidney Game', Animal Behaviour, 42, pp. 1034–5.

Richerson, P. J. and Boyd, R. [2005]: Not By Genes Alone: How Culture Transformed Human Evolution, Chicago, IL: University of Chicago Press.

Savage, L. J. [1954]: The Foundations of Statistics, New York: John Wiley & Sons.

Sterelny, K. [2012]: The Evolved Apprentice: How Evolution Made Humans Unique, Cambridge, MA: MIT Press.