Agents and Goals in Evolution, by Samir Okasha.
Oxford: Oxford University Press, 2018. Pp xiv + 254.

Sometimes, watching ants, it's hard not to feel a sense of pathos. There is a species in Brazil, *Forelius pusillus,* that takes the defence of its nest unusually seriously. To conceal and protect the nest at the end of each day, some of the workers seal off the entrance—from the outside. Left out in the cold night-time temperatures, these ants will never see the morning. But their sacrifice increases the chance that their sisters will.

Faced with an example like this (from Tofilski et al. 2008), we feel an irresistible temptation to describe the situation in *agential* terms. We impute goals, strategies, reasons and interests to the ants. We say that they seal off the nest *in order to* protect their relatives. We say that they sacrifice their own survival *for the sake of* others. Some of the things we are inclined to say may well be anthropomorphic and unjustified by the biological facts. The ants don't wipe away a tear as they leave the nest; they don't wistfully remember the good times. Yet it is far from clear that scientists are wrong to invoke agential concepts like goals, strategies, reasons and interests in serious explanations of the ants' behaviour. After all, these ants really are akin to agents in certain important respects. Their behaviours really are in some sense strategic, flexible, goal-directed, and attuned in agent-like ways to the facts of their situation. There is no obvious way to mark the point at which soppy anthropomorphism stops and accurate description begins.

Examples like these lead to a set of foundational questions about the nature, validity and value of agential thinking in biology. In *Agents and Goals in Evolution*, Samir Okasha is particularly concerned with four such questions:

i. Is there a serious scientific (as opposed to intuitive) rationale for agential thinking?
ii. If yes, what goal should we expect animals to act as if trying to achieve?
iii. On the whole, should we expect evolutionarily optimal behaviours to conform to norms of rational choice?
iv. If so, are there circumstances in which evolutionary optimality and rationality can part ways, such that the optimal behaviour is in some sense irrational?

Okasha tackles these questions with exceptional precision and attention to detail, offering significant new insights regarding all of them. The book is a model of what philosophy of science could and should be: a careful, rigorous, uncompromising search for answers to hard and important foundational questions. It's uncompromising in the sense that, where necessary, Okasha does not shy away from close engagement with the mathematical details of population genetics and rational choice theory at a level of difficulty that may challenge readers who have no familiarity with either field (it seems, to me at least, somewhat more demanding than his *Evolution and the Levels of Selection*, 2006). That said, it's hard to imagine a more effective entry point to some of these debates for those readers who want to get up to speed with the relevant mathematics.

The book is split into three parts. Part I deals with the rationale for agential thinking, Part II with the issue of what, if anything, is maximized in evolution, and Part III with the evolution-rationality connection. I'll focus here on Parts I and III. I'll zoom in on the question of whether the rationale for agential thinking is the same for genes as it is for whole organisms: Okasha suggests it is, whereas I suspect it is not. I'll then turn to the question of whether evolution and rationality part ways in cases involving strategic behaviour in the face of risk. I'll suggest that there is a useful role for "gene's eye" thinking in explaining why they part ways, when they do.

1. The rationale for agential thinking in organisms and genes

One of the most obvious, but also most puzzling, aspects of agential thinking in biology is that biologists apply it to radically different kinds of entities: organisms, viruses, genes, social groups, entire ecosystems and species, or even natural selection itself. Some of these entities are parts or groups of living systems rather than living systems in their own right. Natural selection, of course, is not even an entity; it is a process, albeit one that is sometimes described in personified terms.

Is agential thinking equally legitimate in all these cases? Okasha argues, plausibly, that it is not. He is sceptical of agential thinking applied to natural selection itself: natural selection is not an agent, it is not goal-directed, and it does not reliably maximize any quantity except under restrictive assumptions (see Chapters 1, 3 and 4). He is also sceptical of agential thinking applied to biological groups, except in those rare cases in which the suppression of within-group selection leads the group to display apparent unity of purpose (see Chapter 2). Multicellular organisms—highly integrated groups of cells, like us—can legitimately be regarded as biological group agents, and advanced eusocial insect societies may belong in the same category, but most groups of animals do not. Genes and organisms, however, can both be usefully regarded as agents under a wide range of conditions.

This leads to the question: is the rationale for thinking of organisms as agents *the same* as the rationale for thinking of genes as agents? Okasha claims it is, at least in the context of evolutionary biology (p. 47). In both cases, he suggests, the fundamental rationale is adaptedness: it's because both organisms and genes can be bearers of adaptations, and because these adaptations appear to conduce towards a *single, unified purpose*—reproduction or inclusive fitness in the case of organisms, replication in the case of genes—that agential thinking is justified.

In the case of organisms, there are also *supplementary* rationales which don't apply to genes: an organism's behaviour is often flexible (the behaviour performed is selected from a wide repertoire of possible behaviours) and "goal-directed" in a non-evolutionary, cybernetic sense (the behaviour converges robustly on an end-point from a variety of initial conditions and in spite of perturbations) as well as an evolutionary sense. However, these supplementary rationales are independent of the adaptedness rationale and don't necessarily apply to all organisms, because they don't apply to organisms that are entirely sessile and entirely without flexible or goal-directed behaviour (I'm not sure there are any such organisms—Okasha gives the example of a cactus but immediately qualifies it—but we can accept the logical possibility of such organisms). In the case of genes, Okasha argues, the adaptedness rationale still applies, but the supplementary rationales do not.

The difficulty here is that a rationale that applies equally to both organisms and genes must inevitably be quite thin. The apparent consequence is that whenever we find an entity with parts that seem to conduce towards a common purpose, agential thinking will be useful, even if there is no flexible or goal-directed behaviour. But note that agential thinking is typically neither intuitive nor useful in the case of artefacts. A watch has multiple interacting parts that conduce towards a common purpose, but this does not provide any rationale for regarding the watch as an *agent* rationally pursuing a goal. We think of the watch as a product of design, but not as an agent. We need to say something about what distinguishes organisms and genes from watches, such that organisms and genes invite agential modes of explanation whereas watches invite only artefactual, design-based modes.

I doubt there is a common factor, shared by genes and organisms but not by watches, that would explain the difference. It's more plausible, I think, that different rationales justify agential thinking in the case of organisms and the case of genes. In the former case, I suspect it's the *conjunction* of flexible behaviour and adaptedness that makes agential thinking attractive. It's because organisms behave flexibly that it makes sense to regard them as facing *choices between options*, evaluable in principle as more or less rational in regard to their interests. There is a genuinely choice-like phenomenon here, whether or not it involves choice in anything like the human sense. Evolutionary considerations then supply the relevant evaluation criterion (typically, inclusive fitness).

By contrast, there is nothing literally choice-like about what a gene does at the molecular level. As Okasha notes, individual gene-tokens—particular sequences of DNA inside a particular cell—have little flexibility. They may or may not be expressed, but the gene-token itself does not control this: its expression is controlled by mechanisms of gene regulation. "Gene's eye" thinking, although occasionally done at the level of individual gene-tokens, is more commonly done at a more abstract level, centred on an abstract entity: a "locus" in the genome, at which various different possible forms of a gene—the gene's "alleles"—compete for representation in a population.

Agential thinking about genes and their alleles strikes me as a much more richly *fictional* activity than agential thinking at the organism level. It can be heuristically valuable to imagine a "gene"—here used to refer to a genomic locus, not a concrete gene-token—as if it were faced with a choice among the various alleles it could adopt. Should a gene choose an allele that will promote the survival and reproduction of its bearer, or should it choose a different allele that will promote its own replication at its bearer's expense? The gene's "choice" is just an abstraction from gene frequency dynamics: what really happens is that one allele spreads at the expense of the others. There is more fiction in this way of thinking than in its organism-level equivalent, because there is nothing genuinely choice-like about what is happening. Individual gene-tokens do not literally try on and discard different alleles like pieces of clothing. But this is a sometimes-useful abstraction from a process in which alleles compete with and displace others, leading those which best promote their own replication to become fixed.

So, here is my counter-proposal: a well-adapted, behaviourally flexible organism invites agential thinking because it genuinely faces choices between options, in something like the way we do. By contrast, a gene invites agential thinking because it's helpful to take a complicated, population-level process we struggle to visualize—alleles changing frequency—and reimagine it as a simpler process that is easy to visualize: a gene choosing its allele. The two rationales are distinct. A watch does not invite agential thinking because, even though it still displays adaptedness and unity of purpose, neither rationale applies.

2. Risk and rationality

The difference between organism- and gene-level perspectives resurfaces in the discussion of risk and rationality in Chapters 7 and 8, which is perhaps the part of the book with the widest philosophical interest. Okasha spends most of Chapter 7 arguing that many purported examples of evolution leading to irrational behaviour are unconvincing, because they can be dispelled by choosing a suitable utility function (e.g. a utility function that counts harms and benefits to relatives, or that incorporates inequity aversion), by re-describing the options more carefully, or by refining the norm of rationality allegedly violated. But in Section 7.6 and Chapter 8, Okasha argues that there really can be a parting of ways between evolution and rationality. Here, Okasha discusses models in which natural selection will reliably lead to organisms with irrational preferences regarding risk. Their preferences are irrational in the sense that they cannot possibly be reconciled with standard rational choice theory, regardless of the choice of utility function.

The source of irrationality here is a distinction between two types of risk: *idiosyncratic* and *aggregate* (or systematic). These terms can be applied whenever we have a cohort of agents (in this case, organisms of a particular genotype) in a stochastic environment. Idiosyncratic risks are those such that each individual organism gets its own toss of the coin, and they are such that (by definition) they average out across the cohort. They cause some individuals to do well and some to do badly, but they have no effect on the success of the cohort as a whole. By contrast, aggregate risks are such that the entire cohort gets a single toss of the coin. These risks do not average out. They are things like dry summers or harsh winters. They cause the whole cohort to do well or to do badly.

Crucially, natural selection will tend to favour those alleles that are less vulnerable to aggregate risks. Idiosyncratic risks don't make a difference to the evolutionary fate of an allele. In giving an intuitive explanation of this, I can't help but fall back on the gene's eye view: genes don't need to worry about idiosyncratic risks, because idiosyncratic risks, by definition, average out over the cohort of bearers of the same allele. Genes don't care about the risks that wash out; they care only about the risks that affect the fortunes of their entire cohort of bearers.

This is a familiar idea in finance, where the terms "idiosyncratic" and "aggregate" risk originate. Investors who put money in a fund with broad exposure to a whole index don't need to worry about risks that are idiosyncratic to a specific company. They need only worry about risks that affect the entire index. An allele with a large number of bearers is in a comparable situation. Selection favours alleles, not individual organisms, and in unpredictable environments it tends to favour, over the long run, the ones that are less vulnerable to having their growth rate periodically knocked back by aggregate risks.

This is captured formally in the "geometric mean principle": in stochastic environments, natural selection favours those alleles that maximise the geometric mean (over possible environments) of their bearers' per-capita reproductive output. The geometric mean is a formal representation of risk aversion: compared with an arithmetic mean, it gives more significance to low values, and an entry of 0 for any possible environment entails a geometric mean of 0. But only aggregate risks—risks that affect all bearers and are visible in the per-capita reproductive output of the allele—are relevant here.

As Okasha explains, this means that, if there is a way for an organism to *convert* aggregate risk into idiosyncratic risk through a behavioural choice, alleles that cause their bearers to do the conversion will sometimes be selected even if choosing the conversion would be irrational from the point of view of an individual organism trying to maximize its fitness. Perhaps the most dramatic example (originally due to Robson 1996) involves a stochastically dominated lottery. Suppose an organism must choose between two lotteries: Lottery A, which yields either 1 or 9 offspring with equal probability, and Lottery B, which yields either 1 or 8 offspring with equal probability. Consider a population containing two types of organism: one type (the A-type) has an allele that reliably causes it to choose Lottery A, whereas the other type (the B-type) has an allele that reliably causes it to choose Lottery B. Which allele will be favoured by selection?

Lottery B is stochastically dominated by Lottery A, so it is irrational for any organism to choose Lottery B. But now suppose that Lottery B involves purely *idiosyncratic* risk (every B-type gets its own flip of the coin, and the mean reproductive output of B-types is always about 4.5 in every generation), whereas Lottery A involves purely *aggregate* risk: a single coin is flipped, and the whole cohort of A-types get 9 offspring each, or else they all get 1 offspring each. The geometric mean principle tells us to compare 4.5 to 3, since there is no aggregate risk for Lottery B and 3 is the geometric mean of 1 and 9. In other words, the B-types will be favoured by selection over the A-types. Intuitively, this is because in each generation there is a 0.5 probability that the A-types will experience a collective disaster, and this hurts the long-run growth rate of A-types. By contrast, the B-types are able to maintain a steady rate of growth through good years and bad.

There are various questions one might ask about cases like this. One question is: *should we think of organisms or genes as the agents here?* To my mind, this situation is aptly described as a subtle conflict of interest between the individual organism and the gene. To pursue the financial analogy: imagine an influential investor who, fearful of a looming aggregate risk that will affect the whole country, persuades several companies to move their operations abroad, even though this is costly and risky for each company and merely converts aggregate risk to idiosyncratic risk. We'd naturally describe this as a conflict of interest between the investor and the individual companies. If an allele causes its bearers to pay a cost in their expected number of offspring to convert aggregate risk to idiosyncratic risk, there is a conflict of interest between the gene and its individual bearers for the same basic reason. If this description of the situation is reasonable, then this class of examples can be assimilated to the broader class of cases (discussed in Chapter 2) in which the apparent unity of purpose of the organism is undermined by a conflict between its own interests and those of its genes.

This pertains to our earlier discussion of agential thinking and the gene's eye view. I suggested earlier than gene's eye thinking usually involves taking the gene *qua* genomic locus, not a particular gene-token, as the focal agent. We picture a gene "choosing its allele" from a range of possible variants as if it were an investor, trying to pick the allele that will maximize its representation in future populations. This is an example in which this abstract way of thinking helps us grasp an idea that would otherwise be harder to grasp. The gene chooses the allele that converts aggregate to idiosyncratic risk, even though an organism behaving rationally would choose differently. I also suggested that gene's eye thinking is more richly fictional than organism-level agential thinking. In cases where the interests of the gene and the organism come into conflict—sometimes, as in this case, for very subtle reasons—we have a more elaborate fiction still. Organisms, which are concrete and agent-like, if not always literally agential, share the stage with genomic loci, which are abstract entities and fictitious agents. The overall picture resembles an historical novel in which real figures share the stage with authorial inventions.

Another question one might ask is: *can the trade-off between aggregate and idiosyncratic risk shed any light on the evolution of irrational preferences in humans and other animals?* Okasha says surprisingly little about this. Although he has written on related topics in the past (Okasha 2007), he steers clear of speculation about empirical cases, human or otherwise, throughout this book. The focus here is strictly on theoretical and conceptual issues; readers are left to add their own speculations.

However, as Okasha notes, the connection between the aggregate/idiosyncratic distinction and human behaviour has been explored elsewhere by Arthur Robson and Larry Samuelson (2012), who highlight the importance of *control* to human judgements about risk. Curiously, we spend more time worrying about risks beyond our control, such as disease epidemics and air accidents, than about risks partially within our control, such as risks relating to driving and diet. They speculate that control is a psychological proxy for idiosyncrasy: we have evolved to worry more about uncontrolled risks because they tend to be aggregate, and to worry less about controlled risks because they tend to be idiosyncratic.

Okasha says nothing either for or against these speculations. But he does point out (in Section 8.5) that, once a wedge has been pushed between evolution and rationality, it becomes possible, at least in theory, to use this wedge to generate various different types of irrational behaviour. In any scenario in which a particular pattern of preferences over lotteries would be irrational, we can construct a model in which the selection pressure to convert aggregate risk to idiosyncratic risk causes a population to adopt those preferences. For example, as Okasha cleverly shows (on pp. 218-20), we can exploit this pressure to create a model in which the population evolves to an equilibrium that constitutes an Allais paradox: at equilibrium, the organisms' preferences in one pair of lotteries are irrationally reversed when an equal payoff is added to both lotteries, because the two pairs differ with respect to their imposition of idiosyncratic and aggregate risk.

Allais preferences violate the independence axiom of rational choice theory. Do we have here the beginnings of evolutionary explanation of violations of the independence axiom? Perhaps, but only if the source of the irrationality is that the agent is responding to some sort of proxy for idiosyncratic as opposed to aggregate risk, as in Robson and Samuelson's hypothesis. Is that ever the case? How often do animals actually face choices that give them a degree of influence over the type of risk they face, and how do they choose when they do? This is the point at which theory gives out, and some experimental evidence is needed. The idea is tantalizing, and it strikes me as deserving of empirical investigation. This could be the start of a fascinating research programme, if biologists and social scientists are willing to pick up the baton.

3. Conclusion

I have barely scratched the surface here of the many subtle, rich and illuminating points made in this book. Anyone with a serious interest in the foundations of evolutionary theory and the nature of evolutionary explanation will get a lot out of it, regardless of their disciplinary background.

JONATHAN BIRCH
London School of Economics and Political Science
j.birch2@lse.ac.uk

References

Okasha, Samir 2006: *Evolution and the Levels of Selection* (Oxford: Oxford University Press).

----- 2007, 'Rational Choice, Risk Aversion, and Evolution', in *The Journal of Philosophy*, 104.

Robson, Arthur J. 2006, 'A Biological Basis for Expected and Non-Expected Utility', in *Journal of Economic Theory*, 68.

Robson, Arthur J. and Larry Samuelson 2011, 'The Evolutionary Foundations of Preferences', in Jess Benhabib, Alberto Bislin and Matthew O. Jackson (eds.), *Handbook of Social Economics* (Amsterdam: North-Holland).

Tofilski, Adam, Margaret J. Couvillon, Sophie E. F. Evison, Heikki Helanterä, Elva J. H. Robinson, and Francis L. W. Ratnieks 2007, 'Preemptive Defensive Self-Sacrifice by Ant Workers', in *The American Naturalist*, 172.