Title: Samir Okasha's Agents and Goals in Evolution, a Review

Name: Cailin O'Connor
Affiliation: Department of Logic and Philosophy of Science, University of California, Irvine
Email: cailino@uci.edu

Acknowledgments: Many thanks to Michelle Pham and Samir Okasha for comments on this manuscript.

Samir Okasha's *Agents and Goals in Evolution* considers agential thinking in biology. The book is exceptionally well written. It is thorough, precise, and draws appropriately restrained claims about when and whether agential thinking can be useful. As Okasha points out, such thinking is commonplace, and falls roughly into two camps. First, biologists often conceptualize organisms as agents who seek to accomplish certain goals, and especially as rational agents who seek to promote their own fitness (agential thinking 1). Second, biologists sometimes conceptualize the process of evolution itself as driven by an agent with goals (agential thinking 2). While Okasha largely rejects agential thinking (2), he argues that agential thinking (1) can play a legitimate role in theorizing. This review will describe the progress of the book, and conclude by discussing a topic that attenuates the appropriateness of rational agent thinking in biology.

Okasha begins part I by arguing that treating biological entities as agents is most useful when these entities display a unity of purpose, derived from shared biological fate. In such cases, the entity will typically engage in behaviors that appear instrumentally rational, and coherently lead towards certain goals. Agential thinking can elucidate how these evolved traits and behaviors fit together to fulfill these goals. In addition, as Okasha points out, agents with the proper unity of purpose can often be represented, using decision theory, as *rational* agents whose goal is to promote their own fitness.

A natural question arises: are entities other than organisms – namely genes and groups – usefully thought of as agent-like? Okasha argues that this is most appropriate in understanding traits that otherwise seem to make no sense, like "outlaw genes" that promote their own fitness at the expense of organismic fitness. In these cases, the gene does not share the organism's unity of purpose, and so is usefully treated as a separate agent working towards its own goals. What about groups? As Okasha points out, true unity of purpose is relatively rare in biological groups, but not unheard of. In such cases groups can be treated as agents.

In part II of the book, Okasha moves on to ask: does evolution maximize population fitness? And: does it optimize the fitness of individual agents? As Okasha points out, there are conceptual connections here to agential thinking of both types – if evolution maximizes population fitness, it is arguably acting like a rational agent (agential thinking 2). If organisms evolve to be optimal, then we expect them to be rational as well (agential thinking 1).

This is the most technical part of the book. Okasha goes through several extant arguments for maximization/optimization. With regards to Wright's *adaptive landscapes* (see Wright (1932)) – while evolution will supposedly climb hills in these landscapes, this sort of hill climbing actually only occurs under restricted conditions. This case of agential reasoning 2 is thus a misleading one. Okasha also addresses Alen Grafen's maximizing agent analogy (Grafen, 2014) which suggests that individual fitness maximization is the expected result of natural selection. However, Grafen focuses on frequency independent selection, i.e., where fitness of a genotype does not depend on the prevalence of that genotype in the population. Under frequency dependent selection both average population fitness and individual adaptedness can decrease as evolution progresses.

Okasha then turns to maximization/optimization arguments in *social* evolution. The evolution of altruism looks like a serious problem for optimization (and for conceptualizing of organisms as rational); because altruists (irrationally) decrease their fitness to benefit group members. However, the metaphor can be saved by appeal to *inclusive fitness,* which tracks the offspring an individual causes (rather than direct descendants). Under certain conditions, evolved altruists will behave like rational agents who maximize inclusive fitness. Under other conditions, though, things are more subtle, meaning that the rational agent metaphor cannot always be saved.

Okasha, however, does not take the failures of these theoretical arguments for optimization to negate the usefulness of agential thinking. As Okasha points out, *empirically* we do tend to see adaptation in biology. This, then, is the proper justification for adaptationist approaches and agential thinking. An upshot is that we should be critical of agential thinking (2) – mother nature does not seem to "choose rationally" – but agential thinking (1) is still reasonable in cases where we can see empirically that organisms are well-adapted and act to maximize their own fitness.

Part III of the book addresses the connection between evolution and rationality. Okasha differentiates two important issues here. First, related to part I, rationality *concepts* are useful for thinking about organisms' behavior. Second, *actual* rationality is an evolved trait. As he points out, these are linked by the idea that the behavioral plasticity of many organisms is proto-rational. This fact helps elucidate both how we evolved rational thinking, and also why we can usefully treat such animals as rational agents. In other words, this conceptual link both helps justify agential thinking 1, and explains why this thinking is so often useful.

Okasha starts this part with the question of how and whether rationality evolved. Flexible behavior and psychological states of belief and desire, i.e., proto-rationality, arguably evolved to help organisms deal with variable and complex environments. The next question is whether evolution drives organisms towards more full-bodied rationality. Clearly creatures that have accurate beliefs and consistent desires related to maximizing biological fitness will often tend to do well in the world. But this observation does not mean that evolution will *always* select for full rationality.

Okasha goes through a number of arguments showing how evolved behavior can depart from economic-style rationality, i.e. from utility maximization. First, in evolutionary game theoretic models of the prisoner's dilemma and the ultimatum game, behavior that does not look rational can evolve, i.e., altruism and retaliation. But we need not think of this behavior as irrational if organisms have evolved to value things other than fitness, such as the well-being of others, and fairness. However, if we make this move we need to be careful about treating organisms as rational *fitness* maximizers, which theorists often do. Further cases regard the evolution of intransitive choices, irrational risk preferences, and irrational payoff discounting. In each case, irrational looking behaviors 1) can evolve in models and 2) are empirically observed. In each case, though, Okasha again points out that if we re-contextualize the choice scenario, the behavior will instead seem rational. So again, this means that while conceptualizing organisms as rational fitness maximizers must be done carefully, this conceptualization can still be useful.

Despite his persistent caution, Okasha does not do much to address a ubiquitous set of cases that might be added to this part of the book: ones where constraints related to cognition and evolution necessarily prevent rational behavior. For instance, it might be extremely costly to develop the cognitive apparatus to engage in high rationality behavior. And doing so might require an organism to give up on other desirable cognitive features. It seems to me that many of

these constraints suggest important further limitations on the use of agential concepts in biology. Furthermore, ignoring these constraints has led biologists wrong in real cases.

To give one example from my work, Maynard-Smith (1982) argues that evolution will select for learning that leads to the play of evolutionarily stable strategies (ESSes) in strategic contexts. ESSes make up a subset of predicted rational behavior in the relevant scenarios. His argument is essentially that because ESSes are equilibria in games, other behaviors do poorly against them, and will be selected against. However, learning strategies that quickly adopt decently good behavior often do not allow organisms to learn ESSes. Consider learning generalization, whereby organisms apply learned lessons to novel, but perceptually similar, scenarios. This kind of learning is necessary for successful behavior (since organisms very rarely find themselves in the *exact* same scenarios twice). It also leads to actions that are not perfectly tuned for the exact scenario they are employed in, and thus are not ESSes. But we can use evolutionary models to show that because the speed of learning matters so much to organism payoffs, imprecise, quick, generalizing strategies do, in fact, evolve (O'Connor, 2017). In other words, ignoring cognitive trade-offs led Maynard-Smith to incorrectly predict the emergence of rational behavior.

I do not think this sort of case is a serious problem for Okasha. As described, throughout the manuscript he advises caution in using (rational) agent concepts in biology. Furthermore, as noted, he urges theorists to use empirical work, rather than theoretical arguments, in deciding when agential thinking is appropriate. I think there is a more specific take-away he might have emphasized that seems right given the limitations he focuses on, and the ones he does not, for agential thinking. While we can often treat organisms as rational agents for descriptive purposes, we should not do so when trying to predict behavior. There are too many reasons why evolution may not have led to straightforwardly rational behavior in any novel case, even if in many cases we can observe that it did.

References:

Grafen, A. 2014. The formal darwinism project in outline. *Biology & Philosophy*, *29*(2): 155-174.

Maynard-Smith, J. 1982. *Evolution and the Theory of Games*. Cambridge, UK: Cambridge University Press.

O'Connor, C. 2017. Evolving to generalize: Trading precision for speed. *The British Journal for the Philosophy of Science*. *68*(2): 389-410.

Okasha, S. 2018. *Agents and goals in evolution*. Oxford, UK: Oxford University Press.

Wright, W. 1932. The roles of mutation, inbreeding, crossbreeding, and selection in evolution. *Proceedings of the International Congress of Genetics*. 1:356-366