

## Evolution and Rational Agency

by

Geoffrey Sayre-McCord  
(UNC/Chapel Hill)Introduction

*Morality requires what nature does not permit.* That is the common conclusion of a range of familiar arguments to the effect that morality is a chimera or at best a useful fiction. Thus, for instance, some say that the claims of morality presuppose the existence of sui generis, intrinsically motivating, properties of a sort nature excludes. Alternatively, some say that moral truths, were there such, would have to be both synthetic and a priori in just the way a proper understanding of nature shows to be impossible. Still others observe that morality sometimes requires and in any case regularly valorizes an altruism with which the forces of natural selection would quickly dispense. And, finally, some maintain that morality would apply to us only if we possessed a robust kind of rational agency that no clear-eye view would suppose we, natural creatures (and products) that we are, could have.

These are, as I say, familiar arguments. They each play out against some characterization or other of what morality requires and then proceed to maintain that what is required is not to be found in our world. The first two arguments, it seems to me, are wrong about what morality requires, though they are right that if its requirements were as they suppose, the requirements would not be met. The second two arguments, in contrast, seem to me to get right what morality requires. If they fail, and I think they do, it is because there is in fact room in our world for the altruism and rational agency morality requires. In what follows, I will leave aside the first two arguments altogether, I will pause on the third argument concerning altruism, and will then concentrate on the fourth concerning rational agency.

It's worth noting that, when it comes to the last two arguments, the questions aren't just whether altruism and rational agency are compatible in principle with nature, nor even whether they might themselves be natural traits or capacities, but whether, in addition, given what we understand of nature -- and especially evolution and natural selection -- it is reasonable to see altruism and rational agency (of the sort morality requires) as products of nature. Showing this in a compelling way involves showing that they plausibly emerged and have, in fact, persisted in a context shaped by natural selection. To think they did not is to think that what morality (putatively) requires nature hasn't actually allowed (or that nature's resistance notwithstanding, something un- or extra-natural has made what morality requires possible).

To hold that some trait or capacity has emerged and persisted in a context shaped by natural selection does not require thinking that the trait or capacity has itself been selected for. Traits emerge in the first place thanks to mutation, random drift and changes in population and then stand or fall, in the circumstances, thanks to their relative success in reproducing. A trait's success depends on its successful replication. To be durable, new traits, whatever their source, must survive in the face of selective forces. Even then, of course, a trait that is not itself evolutionarily advantageous may survive, if either it piggy backs on some other trait that is sufficiently advantageous or it fits together with other traits that, taken in combination, prove advantageous in comparison with the alternatives.

With this in mind, the most satisfying way to establish that nature allows the altruism and rational agency morality requires would be to show, first, that they would likely have emerged as available traits and, second, that once on the scene they would have proven to be robust enough to survive. Those of us who believe the traits in question are not only possible but often enough actual of course might feel comfortable skipping both steps since their existence establishes that the traits did happen to emerge and their persistence establishes that they are sufficiently robust to endure selective pressures. As it happens, a fair number of people working in ethics take just this attitude. Whatever morality requires nature must allow (or at least not successfully resist) since morality is neither myth nor madness. I think this is the wrong road to take for at least two reasons. First, skepticism about morality is worth taking seriously. Second, understanding the ways in which altruism and rational agency might fit with evolution offers a valuable control on what we might reasonably take to be actual.

Altruism

A significant proportion of the literature on evolution and ethics has focused on altruism. The reason for this is fairly straightforward: a lot of people have supposed that genuine altruism, to the extent that it regularly involves sacrifice for the sake of others, could not possibly survive as a trait in the face of natural selection. After all, even if, through mutation, random drift, or population change, a community of altruists were to emerge, they would rather quickly find themselves infiltrated and then overrun by those who would benefit from the altruism of others without sacrificing themselves. As long as the altruistic and exploitative traits (or the characteristics that underlay them in context) are heritable, it seems, the latter will predictably have more progeny, who have more progeny... on down the line, each time passing on their evolutionarily advantageous trait. The apparently inevitable upshot is that altruism, while it might perhaps pop up here or there as a result of random mutations, is not sufficiently robust to stand in the face of competition.

Embracing this line of thought, many have supposed that the truth of evolutionary theory establishes that at bottom people must actually not be altruists, whatever appearances they might keep up to others or themselves. Nature just won't allow otherwise.

There's a lot to be said about this argument that shows it goes by much too quickly. And a good deal that is subtle and important has been said. Without belaboring the issues it is worth noting several things.

First of all, the argument (as I have sketched it) elides an important distinction between what we might call individual advantage (or disadvantage) and evolutionary advantage (or disadvantage). As I will be using the terms, behavior counts as individually advantageous if it contributes positively to an individual's well-being (whether that is a matter of satisfying her preferences or acquiring something of value or something else); whereas behavior counts as evolutionarily advantageous if it contributes positively to an individual's fitness (where fitness is not a matter of well-being but of successful replication).<sup>1</sup> Various behaviors might be individually advantageous, yet evolutionarily disadvantageous, and vice versa. Acting in a way that contributes to one's well being may get in the way of reproducing, while reproducing may seriously compromise one's own well being.

Similarly, second of all, the argument fails to distinguish between acting with the intention of doing what is advantageous (either individually or evolutionarily, for oneself or another) and doing what is in fact advantageous. This is an especially important distinction when it comes to discussing the altruistic behavior morality sometimes requires and in any case valorizes. For this is (at least roughly) behavior that is done with the intention of doing what one takes to be individually advantageous for another, where the intention is unmediated by the intention to do what is individually advantageous for oneself. Behavior that is done with this unmediated intention may or may not actually have the results intended. And even if it does, those effects may or may not actually be individually advantageous for the person one seeks to help. At the same time, while altruistic behavior takes time and energy away from intentionally pursuing one's own individual advantage, it doesn't at all follow that such behavior is either individually or evolutionarily disadvantageous for the agent.

Thus, it may well be that a community of altruists – individuals who intentionally act to advance the welfare of others, where they do not do so

---

<sup>1</sup> Drawing the distinction in this way allows talk of individual advantage with respect to any thing that might count as better or worse off and allows talk of evolutionary advantage with respect to anything that might successfully replicate. Thus, for instance, it may make sense to talk of both individual and evolutionary advantage when it comes to various social institutions and also when it comes to genes.

merely as a means of advancing their own welfare – might well in fact be acting in ways that are both individually and evolutionarily advantageous. Whether in fact, and not merely in principle, this possibility is realized is of course terribly difficult to establish.

The first question, when trying to establish it, is whether the traits at issue, the various dispositions to act (e.g., with the intention of helping another without the intention of helping oneself), have an appropriately heritable ground or basis. Reasonably impressive evidence that they do is becoming available thanks to recent work in primate ethology while various epidemiological studies (especially those involving twins)<sup>2</sup> seem to evidence an important role of genes in the determination of behavioral repertoires. Of course, to hold that the traits do have a heritable basis is not at all to hold that their transmission and expression is independent of environmental (including social) factors.

Assuming the traits do have an appropriately heritable ground, the next question is whether, and in what ways, they might prove evolutionarily advantageous. In general, there seem to be two ways in which behavior that is individually disadvantageous (whether motivated by an altruistic concern for the welfare of others or not) – even when it proves to be seriously disadvantageous (say, because it leads almost immediately to death) -- will prove nonetheless to be evolutionarily advantageous (at the genetic level).

First, when the behavior enhances the fitness of others who are related to the agent genetically – when there is effective kin selection – the behavior will (in proportion to the closeness of the kinship relation) actually contribute to the replication of the agent's genes. Self-sacrificing behavior governed by kin selection thereby contributes to inclusive fitness (to the propagation of one's genes, whether via direct descendants or not) even though it contributes neither to individual well-being nor to classical fitness as measured by one's own reproductive success and the reproductive success of one's direct decedents.<sup>3</sup>

Second, even in the absence of others who are genetically related, self-sacrificing behavior (again, whether motivated by a concern for the welfare of others or not) will prove evolutionarily advantageous in contexts in which others are disposed to reciprocate, as long as the reciprocation has the effect of enhancing inclusive fitness. If one is in a community of reciprocal altruists

---

<sup>2</sup> Kendler, K., and Prescott, C. "Cannabis use, abuse, and dependence in a population-based sample of female twins" *American Journal of Psychiatry* 155(8):1016-1022, 1998; "Cocaine use, abuse, and dependence in a population-based sample of female twins," *British Journal of Psychiatry* 173:345-350, 1998; "Genetic and environmental contributions to alcohol abuse and dependence in a population-based sample of male twins," *American Journal of Psychiatry*, Vol. 156, No. 1, January 1999, pp. 34-40.

<sup>3</sup> Hamilton, 1964 The evolution of altruistic behavior. *The American Naturalist*, 97, 354-6.

(who are good at detecting whether others are likewise disposed) being a reciprocal altruist oneself will be evolutionarily advantageous even when the disposition leads one to sacrifice one's own welfare.

Kin selection and reciprocal altruism both work to make individually costly behavior evolutionarily advantageous. There's some evidence that, of the two, kin selection is a greater influence and better predicts patterns of cooperation and restraint.<sup>4</sup> Yet there is also substantial evidence that reciprocal altruism takes hold in populations along lines that are independent of relatedness.<sup>5</sup> The point here, however, is just that the apparent tension between natural selection (especially as it is sensitive to inclusive and not just classical fitness) and a disposition to sacrifice oneself for others is not nearly as serious as one might, and many have, thought. Thus there is reason to think that once altruism emerges as an available trait there are realistic circumstances under which the trait would prove to be adaptive (i.e. evolutionarily advantageous) even when it led those possessing it to sacrifice their own welfare for the welfare (or indeed the evolutionary advantage) of others.

In paradigm cases, a trait proves evolutionarily advantageous thanks to its direct contribution to the longevity of its possessor and so indirectly to the possessor's fitness. Kin selection and reciprocal altruism both, to the extent they work by increasing inclusive rather than classical fitness, diverge from the paradigm. Yet they maintain a focus on the individual and her traits. In contrast to this, many think that natural selection applies to groups as well as to individuals.

Darwin himself embraced this idea:

A tribe including many members who, from possessing in high degree the spirit of patriotism, fidelity, obedience, courage and sympathy, were always ready to aid one another, and to sacrifice themselves for the common good, would be victorious over most other tribes.

--Charles Darwin, *The Descent of Man* (1871)

The idea is that traits that groups share (thanks to the behavior of their members) may prove to be differentially advantageous to the groups in ways

<sup>4</sup> Betzig, Laura and Paul Turke: Food sharing on Ifaluk. *Current Anthropology*, 1986: 397-400. But see Kaplan, Hillard, and Kim Hill. 1985. Food sharing among Ache foragers: Tests of explanatory hypotheses. *Current Anthropology* 26: 223-246 for a study that suggests that intra group sharing did not follow kinship lines.

<sup>5</sup> See Wilkinson 1990 for a discussion of cooperation among vampire bats, Heinrich and Marzluff 1995 for a discussion of sharing among ravens, and de Waal (1989 and 1997 "The Chimpanzee's Service Economy: Food for Grooming," *Evolution and Human Behavior* 18, pp. 375-386) for a discussion of chimpanzees.

that then confer selective advantages on the members of the group, and so on the traits they have.<sup>6</sup> Thus if one group fares better than others with which it competes because those within it are disposed (under appropriate circumstances) to sacrifice themselves for others in the group, members of the group will on average enjoy some evolutionary advantage thanks to their disposition. Although this is widely acknowledged as possible, there's substantial disagreement as to how significant group selection actually is. It comes into play only under the very special circumstances (only when, for instance, discreet and comparatively long-lived groups compete without intermingling, and more robustly when there are reliable methods for identifying and excluding those with traits that are individually advantageous to them but disadvantageous to the group) and even when it comes into play the selective pressure it exerts is predictably met by strong and often opposite pressures generated by individual selection due to intra-group competition. In any case, to the extent group selection does come into play, it obviously might favor altruistic behavior, especially as the disposition is discriminating as between those who are and those who are not members of one's own group.

Discussions of altruism, in the context of evolutionary theory, tend to focus not on what morality recognizes as morally important – acting in response to the perceived need of others with the aim of helping them – but on behavior identified without regard to its proximal causes, and specifically without regard to whether it is motivated by a concern for the welfare of others.<sup>7</sup> This is especially true of the work that focuses on the apparent precursors of morality to be found among those who are not primates.

There is work on altruism, however, that does try to get a handle specifically on behavior motivated by the recognition of another's need. And, as De Waal notes, when it comes to getting a handle on morality "We are concerned here with motivations and intentions. Regardless of how care is being allotted, the caregiver must be sensitive to the situation of the other, feel an urge to assist, and determine which actions are most appropriate under the circumstances."<sup>8</sup> Work that takes this into account makes a significant advance, from the point of view of understanding morality, over the work on altruism that focuses merely on behavior and its triggers. What this work seems to establish

<sup>6</sup> See *Unto Others*, by Sober and Wilson

<sup>7</sup> Sometimes the focus is specifically on behavior that increases the fitness of others, while simultaneously decreasing the agent's fitness. Even so characterized, "altruistic behavior" may well be given an evolutionary explanation by appeal to kin selection and reciprocal altruism as long as the fitness that is at stake in identifying the behavior is classical (i.e. non inclusive) fitness. If, alternatively, the actions in question are those that do not contribute even to inclusive fitness of the agent, the only evolutionarily viable explanation on offer seems to be that provided by group selection.

<sup>8</sup> *Good Natured*, (Harvard University Press, 1996), p. 63

is that some animals have acquired the disposition to help others, where this disposition is guided by the cognitive capacity to recognize others' situation (which this apparently involves generalizing from their own experiences as well as learning to respond to sub-group sensitive markers of need). Moreover, recent work in primate ethology suggests that in addition to "being nice" in this context sensitive, altruistic, way, some primates have the capacity to regulate their behavior in light of norms that vary from one social-group to another. They are able, for instance, to determine that certain behaviors will be enforced and others punished and able too, in light of that, to conform their behavior to the norms in force.

Needless to say, both capacities – the capacity to recognize and respond sympathetically (and appropriately) to the needs of others in one's group and the capacity to recognize and conform one's behavior to extent norms – are crucial building blocks for morality. It is hard to imagine genuinely moral agents who lacked these capacities.

Nonetheless, those who do have these capacities are not thereby properly seen as moral agents. As Aristotle pointed out, when he was defending his theory of character acquisition, there is an important difference between those who act in some way that virtue would demand, say by doing the just thing, and those who do that act in the way a just person would. Specifically, he thinks the satisfaction of three distinct conditions is necessary. The agent must realize that what she is doing is just so that her acting as justice demands isn't inadvertent. In addition, though, she must be doing the just act for its own sake and not (merely) from some ulterior motive, as a means of achieving some further end. And finally, her being such as to do the just thing because it is just (and not merely because it happens to be a way of avoiding punishment, or impressing others, or staying out of hell) must be such that she has a settled disposition to do just things for their own sake. A person who satisfies the first two conditions but does not yet have settled in to the relevant disposition will count, for Aristotle's purposes, as a person on the way to acquiring the relevant character trait, but not as someone who has made it and so not as someone who does the virtuous thing precisely as a virtuous person would. For our purposes, though, the second condition is what is crucial. An agent is not a moral agent unless she is able not merely to respond to the needs of others and conform her behavior to enforced norms, but also to do what she does because she judges that virtue requires it. She must be able to take justice as providing a reason. As Kant emphasized, famously, there's a difference between merely acting in accord with duty and acting from duty, where the latter requires a distinctive capacity. This capacity, to guide oneself by one's moral judgments, is central to moral agency. And it is a special case of a more general capacity that is I believe at the core of rational agency. I focus on this more general capacity (the

having of which is a necessary condition of moral agency) in the rest of the paper.

### Successive Approximations of Rational Agency

Kant introduces a view of rational agency when he maintains that "Everything in nature acts according to laws. Only a rational being has the power to act according to his conception [representation] of a law, i.e., according to principles..."<sup>9</sup> He immediately goes on to treat the conception of a law as the conception of something as "practically necessary, i.e., as good." In perfectly rational agents such representations are sufficient for determining the will. But when it comes to imperfectly rational agents -- agents who might fail to do what they judge to be practically necessary, that is good -- the representation is of a command or imperative with which the agent might fail to comply.

In order to get a handle on what is distinctive about rational agents, I am going to move quickly, by way of successive approximation, from undifferentiated "everything in nature" towards "rational agency" with three aims in mind. First, I hope to bring out just how sophisticated an agent might be without being a rational agent in the sense that seems to be presupposed by morality. Second, in the process, I hope also to characterize the complexity such not-yet-rational but very sophisticated agents have in a way that makes it plausible both that the metaphysics they would require is naturalistically tractable and that they might reasonably be expected to enjoy an evolutionary advantage in familiar circumstances. And third, I hope that by backing right up against rational agency, by way of these successive approximations, it will be easy to focus well on what finally is necessary for rational agency to come on the scene.

So my approach here will be to identify successive subsets of things in nature. I begin by noting that among the things in nature, some of them represent the world. Thus photos, ideas, paintings, reports in newspapers, signs by the road, as well as humans, represent the world as being a certain way.<sup>10</sup> Among these things, some, but not all, act on the basis of the representations they have, moving or not as a result of how they represent the world as being. We can, for instance, easily imagine building a little robot that has the capacity to represent various features of the world as being one way or another and that has the capacity too to respond differentially depending upon how it takes the

<sup>9</sup> Grounding for the Metaphysics of Morals, Immanuel Kant (Hackett Publishing, 1993), translated by James Ellington, p. 23.

<sup>10</sup> Needless to say, a lot is required in order for something to count as having representations at all, and even more for those representations to be representations of the world as being a certain way. What exactly is required, I won't explore here. I will note, though, that a good variety of things seem to have whatever it takes.

world to be. Similarly, animals regularly seem to respond to their representations of how the world is, moving about in ways that are guided by those representations.

Many such things are simply, as I will put it, *stimulus-response agents*. Some, however, have the capacity not merely to represent things as being a certain way, but also the capacity to represent things as being such that, as a result of their own intervention, they would turn out one way rather than another. Such beings can, in effect, represent different possible courses of action and have the capacity to respond differentially to those representations. They are *planning agents*, able to respond differentially to various prospective courses of action.

Some agents, however, are more than merely planning agents thanks to their capacity to represent other agents as responding differentially to their representations of their own prospective options, where those options are seen by these agents as dependent in part on the actions of others that also represent their prospects as interdependent. Agents that have, in addition to this capacity, the ability to respond differentially to such complex representations, are *strategic agents*. Strategic agents represent not just how things might be as a result of their intervention, but also represent other agents as agents responding to representations of how still others will act in various situations. And strategic agents have the capacity to act on the basis of those representations. Thus how they act depends not just on how they take the world to be, but on how they think the world might be as the result of their own intervention and the intervention of others able likewise to respond to their understanding of their environment and options.

With these agents on board we have, in effect, all that decision and game theory concern themselves with (to the extent they identify an agent's preferences with the patterns of differential response to various prospective options). Assuming that an acceptably naturalistic account of representation (and concept possession) in general can be given, it seems clear that the presence of Strategic Agents in our world introduces nothing metaphysically worrisome. Moreover, it seems clear that having the capacity to represent and thereby respond to the ways the world might be as a result of one's intervention would often be salutary both from the point of view of individual welfare and evolutionary advantage. Indeed, the evidence provided by recent studies of primates makes clear that in fact the relevant capacities have successfully secured a place in nature in a context shaped by natural selection.

Before moving on to Rational Agents, it is worth noting just how sophisticated the Strategic Agents might be absent rational agency. They might well, for instance, have psychological concepts that put them in the position to represent whether and to what degree various options would cause them

pleasure or pain and to represent whether and to what degree those options would cause others pleasure or pain. And they might be disposed either to pursue the prospect of their own pleasure or the pleasure of others. Also, overlaying these possibilities, such agents might introduce rules for behavior to which they are disposed to conform and they might also acquire the disposition to enforce those rules by intentionally causing pain to those who violate them. All of this is possible (and indeed apparently actual) in the absence of a capacity to represent the standards *as good* and the violations *as wrong*.<sup>11</sup>

### Rational Agents

What more is needed, then, in order for an agent as sophisticated as strategic agents might be, to be rational agents as well? The short answer is that in addition to having (i) the capacity to represent how things might be as a result of their intervention and the intervention of others and (ii) the capacity to act on the basis of those representations, they must also (i) be able to represent the different options as better or worse, as more or less worth pursuing and (ii) be able to act on the basis of this evaluative representation. The crucial addition, of course, is the addition of a capacity to represent various options as better or worse. Once that capacity is in place, the capacity to act on the basis of that representation would seem not significantly different in kind from the capacity to act on the basis of other representations – a capacity that has been on the scene from the start with stimulus response agents.

What does it take for an agent to have the capacity to represent various options as better or worse? When does the agent have the concept of value this would require? I think there are two paths to follow in approaching these questions this question and the paths should converge. The first path requires deploying our concept of value in order to determine which of their options are in fact good options for them in their situation. Against that background, the task is to determine whether they are properly responsive to the differences between the options that are, and those that are not, worth taking. Significantly, their being properly responsive is not the same as their reliably taking or reliably tracking, those options that are worth pursuing. On the one hand, an agent may reliably take those options without having the requisite concept at all, say thanks to effective hardwiring. On the other hand, an agent may have the requisite concept and yet regularly, even systematically, get wrong which things are worth doing even as it does actually possess the relevant concept. In order for an agent to count as properly responsive to the value of the options available – in order for the agent to be responsive in a way that constitutes grounds for attributing a concept of value – the agent must respond appropriately to what,

---

<sup>11</sup> De Waal's work is especially intriguing on this front, since he has found strong evidence that communities of primates are able to introduce and enforce various social norms that seem to shape behavior by (at least) altering incentive structures.

given its situation, would be evidence for the value of various options. To this extent, the situation is directly analogous to the one we would be in when trying to determine whether some agent has the concept of blue. For in the case of color what we would need to do is see not whether the agent responds reliably to blue things – it could do that without having a color concept at all – but whether it responds appropriately to what, in its situation, would be evidence for the blueness of various things.

Of course talk of responding appropriately to evidence is horribly vague, not least because whether some consideration or experience counts as evidence or not it is extremely context sensitive, and much of the context to which it is sensitive is the context constituted by the other concepts the agent has available. Not surprisingly it becomes plausible to attribute the concept of blueness to some agent only as it becomes plausible too to attribute a range of other concepts. Similarly, it will be plausible to attribute the concept of value to some agent only as it becomes plausible too to attribute a range of other concepts. Still, the dispositions and sensitivities the presence of which would constitute together (as it would seem) an appropriate responsiveness to available evidence that options are more or less valuable are all of a kind with those that would constitute together an appropriate responsiveness to available evidence that things are blue or not.

The underlying idea, here, is the familiar one that having a particular concept is a matter of being in a certain functional state, albeit one that is (unavoidably) characterized in terms of being sensitive to evidence, to reasons for thinking the concept in question applies. This feature of the characterization means that it does not hold out hope of successfully reducing the concept of value to some evaluatively neutral description of dispositions. All the same, though the characterization seems irreducibly evaluative, the dispositions that would allow the characterization to fit appear not to involve any mysterious metaphysics or occult sensitivities.

The very fact that the crucial dispositions are similar in kind to those that would underwrite ascribing to the agent the concept of blueness, though, raises the worry that the concept the dispositions would underwrite ascribing would not actually be evaluative. Here is a way to think of the worry. Suppose one were inclined to hold, as many have, that our concept of value is a concept of being such as to secure approval under certain circumstances. With that account in hand, the challenge of determining whether an agent has a concept of value becomes the problem of determining whether it is appropriately sensitive to evidence that things would or would not secure approval under certain circumstances. Suppose, then, that it emerges that some agent is sensitive in the appropriate way to evidence that things would secure approval under the relevant conditions. It is at least tempting to think the agent might still just have

a non-evaluative concept of a disposition – the disposition to secure approval – and not a concept of value at all.

What does it take for a concept to be an evaluative concept? This question suggests the second path one might take to determining whether some agent has the capacity to represent various options as better or worse. This second approach starts by deploying not our concept of value (in an effort to determine whether the agent is appropriately responsive to the relative value of her options) but our concept of an evaluative concept. Against the background of having established that the agent is appropriately sensitive to evidence concerning the value of her options, the question is whether the concept there is play (that is differentially applied in response to the available evidence) is an evaluative concept or not. The guiding thought is that the concept, whatever it is a concept of, cannot be a concept of value if it is not an evaluative concept.

Fair enough, I think. But at this point it is difficult to say just what our concept of evaluative concepts requires. A common suggestion, though one that seems inadequate, is that a concept counts as evaluative if it is action guiding. On this view, to see something as good is, in effect, to be attracted by it. The inadequacy of the suggestion comes out clearly with reflection on the various representing agents that fall short of being rational agents. They are all such that certain representations are, for them, action guiding. Yet when the concepts mobilized in those representations succeed in guiding behavior (by prompting the agent to act in various ways) they are not thereby evaluative concepts. For instance, when an agent develops the disposition to avoid red things, the concept of redness hasn't then become an evaluative concept, just a causally efficacious concept. Similarly for agents that are moved to take options that they represent as resulting in pleasure. The representation of future pleasure is in this case causally effective, but that doesn't mean that it is an evaluative concept. What then is required?

A useful way to approach this question, I think, is to go back to the dispositional account of value I mentioned (not because it is especially plausible but because it is helpfully simple and clear). The worry was that there is evidently an important difference between having the concept of a dispositional property (the property of being such as to give rise to approval under certain circumstances) and having the concept of value.

Before exploring this worry, it is worth noting that we might here contrast a concept of a non-evaluative property (say the dispositional property of giving rise to approval) with the concept of an evaluative property (say the property of being approvable, that is, worth approving) or we might contrast a non-evaluative concept of a property with an evaluative concept of a property. Moreover, we might think that in order for a concept to be a concept of an evaluative property the concept we have must itself be an evaluative one.

With that last idea in mind we can turn back to the dispositional account of value. In order for that view to be plausible it seems reasonable to think that the things that secure approval under the specified conditions are such that they *should* secure that approval, that such a response is *good* or *appropriate* or *justified* under the circumstances.<sup>12</sup> Only then would it be plausible to think that the approval secured is of something good. But what is it to ask whether those things that secure approval under the specified circumstances should? According to the dispositional theory on offer, it is in effect to ask whether the fact that those things secure approval under those circumstances would itself secure approval under those circumstances. If the securing of approval under those circumstances would itself secure approval under the appropriate circumstances, then one challenge to the dispositional account would have been met. So long as a person who embraces the dispositional account has independent reason for thinking it plausible, she has the resources to respond to the demand that the things that prompt the approval should merit that approval.

The response, that the approvals themselves secure the appropriate approval has a strong aura of begging the question against those who would advance a different account. Yet two things are significant about it. If the initial demand is legitimate (as I think it is) it limits candidate accounts of value to those that would make applying the account to itself at least intelligible. That the dispositional account does this is not trivial. In addition, the initial demand is met only if, when a proposed account is applied to itself, it in fact lives up to its own standard. Meeting this requirement is also not trivial. It may well be, for instance, that our own patterns of approval would not ratify themselves – that on reflection we would not approve of our approving of the things we do in the way we do. So if the dispositional account survives the test, it accomplishes no small feat.

Right now, though, my interest is not in whether the account meets the demands, but in the relevance of the demand. For it seems to me that a distinctive feature of evaluative concepts is that the standards for their application are always in principle themselves open to evaluation -- and answerable to the results. If a concept is an evaluative concept we can ask not just whether it is being correctly applied given the standard it embodies but can ask as well, of that standard, whether it is a good one, whether we are justified in relying on it. Thus, to turn back for a minute to the dispositional account of value, the standard proposed by the account (as set by what would be approved of under certain conditions) is liable to challenge as perhaps not a good or justifiable one. We can ask legitimately whether it is good or justifiable that it is

<sup>12</sup> This is a point John McDowell makes in “Values and Secondary Properties” when discussing the suggestion that dangerousness should be understood in terms of dispositions to prompt fear.

the standard to be used in determining what is of value. There is an important contrast, here, with non-evaluative concepts (e.g. those of color) that are as they are, we might say, without having to be such that the standards for their application are justified or good. Once we have the concept of blueness up and running, to ask of the standard for its application whether the standard is a good one is to ask not whether we’ve gotten the standard right, but whether we should continue to be concerned with distinguishing between those things that are blue and those that are not. In contrast, once we have an evaluative concept up and running – say the concept of value -- to ask of the standard for its application whether the standard is a good one is to ask whether we have the standard right, it is not to ask whether we should continue to be concerned with whether things are good or not.<sup>13</sup>

This feature of evaluative concepts both reflects and in a sense explains the essential contestability of our evaluative concepts. For to discover of a concept that it is not contestable – that the standards embodied in its deployment are not open to challenge as unjustified – is to discover that the concept is not an evaluative one. The liability to challenge goes hand in hand with the sort of claim to legitimacy that evaluative concepts carry. Thus to discover of some population’s concept that it is not liable to such a challenge, that they would reject as out of place the question of whether the standards in play are good, is to find grounds for thinking the concept they are using is not an evaluative one. To put things in a slightly different way: just as our grounds for attributing various familiar non-evaluative concepts involve discovering whether those who supposedly possess it are appropriately sensitive to evidence for its applicability, so too with evaluative concepts, though in the case of the evaluative concepts the relevant evidence concerns the justifiability of the standards in play.

Thus, for example, if we were to come upon a community that used a term that sounded a lot like “right,” that was applied regularly to things that were, as we see them, actually right, on the grounds that they met some social norm that was in place, we’d reasonably think that they are deploying the same concept we are when we judge of things that they are right. Yet suppose we discover that their use of their term is securely determined by whether or not the things in question accord with the social norms that are in force, regardless of whether they think there is any good justification for those norms. We might discover this in any number of ways. If we did, we would have some grounds

<sup>13</sup> It is worth noting that the test here on offer is not a reflexivity test. The question is not, of each evaluative concept, does it satisfy itself. When it comes to the concept of badness, for instance, which is just as much an evaluative concept as that of goodness, the crucial point is that questions concerning the value of the standard we use in applying the concept of badness are probative with respect to our having the right standard. To discover that the standard is one that implies distinctions we cannot justify is to discover it is not, actually, the right standard for determining what is (and is not) bad.

for suspecting that their concept is not a concept of rightness, and indeed not an evaluative concept at all, but instead a concept that corresponds roughly to our concept of “socially accepted” or “allowed by convention.” Their failure to consider whether the norms in force are justifiable and their resistance to challenges pressing the point would be evidence that they are not sensitive to the evidence relevant to the application of the concept of rightness. Of course, that resistance isn’t decisive, since there might be explanations of the failure and the resistance that is compatible with or even suggests that they are after all using the concept of rightness. This would be the case, for instance, if the resistance reflected a conviction that the challenges were disingenuous and ill motivated or that they were more likely to lead away from the truth that towards it. In such a situation, the resistance to particular challenges might reflect a deeper (though perhaps seriously misguided) concern with being sensitive to the appropriate evidence. So the point isn’t that those deploying evaluative concepts necessarily welcome or respond appropriately to demands for justification. Regularly they do not. Yet if the concept in play is genuinely evaluative my suggestion is that it must be such that reflection on their justification is both appropriate and probative with respect to the proper understanding of their application.

Bringing this all back to the difference between merely strategic agents and rational agents, my suggestion is as follows. For agents to be rational agents they must possess and be appropriately responsive to evaluative concepts that put them in the position to represent available options as (in effect) better or worse. This, in turn, requires that they have a complex set of dispositions that allows them to be appropriately responsive to instances of value, where being appropriately responsive requires not reliable tracking but responsiveness to evidence of the relevant sort. So much we get as a requirement by following the first of the two paths I identified above. When the relevant dispositions are in place we have grounds for thinking a concept is in play such that we can render intelligible a distinction between how things seem and how they are, and so can see the agents in question capable of representing (and misrepresenting) valuable (and valueless) things as being a certain way. But we have grounds for thinking they are representing those things *as valuable* only if, in addition, the concept they are deploying is an evaluative concept. And that means, if I am right about our concept of evaluative concepts, that their deployment of the concept in question must be sensitive not just to whether things satisfy a certain standard but also whether that standard is itself justified (which is to say, at least roughly, that the standard itself is such that it satisfies the standard it sets – and its use as a standard its sensitive to its satisfying this test.

When all of this is in place, when the agents in question have acquired the relevant responsive and reflective dispositions, they have thereby acquired the concept of value. And to the extent they have the ability to respond appropriately to their evaluative judgments of their options, they qualify as

rational agents in the robust sense that seems to be required by morality. Importantly, in so qualifying it seems they do not need to have acquired any traits that implicate naturalistically intractable properties. Moreover, the traits they have acquired, which give them the capacity to respond differentially and reflectively to the value of the options they have seem quite clearly to empower them in ways that at least might be evolutionarily advantageous (for much the same reason other cognitive resources that allow representing and reasoning about the world might prove evolutionarily advantageous). Whether in the end the relevant capacities are metaphysically modest will turn, at least in part, on what the best theory of value ends up saying value is like. If it is peculiar enough, our ability to respond appropriately to it may well presuppose strange, even occult, capacities. But the more a theory suggests that that is what is required, the more we have reason to think there is no such thing to which we might actually respond...

[Points to press: (i) idea that challenges to the standards are probative with respect to our understanding of the concepts in question. To do this well I need to clear up the difference between challenging evidence that criteria are satisfied and challenging the criteria themselves. (ii) idea that this feature of evaluative concepts helps explain the distinctive suitability of the method of reflective equilibrium when it comes to developing an appropriate conception of various evaluative concepts. (iii) idea that the usually secure distinction between a better theory of x, and a theory of a better x, collapses when the concept of the x in question is an evaluative concept. Break paper into two, one on Evolution and Morality, the other on Evaluative Concepts.]