Rational Agency and the Nature of Normative Concepts[1]

Geoffrey Sayre-McCord

[DRAFT, January 23, 2015]

Introduction

Primate ethologists interested in the evolutionary roots of morality have recently discovered evidence that, in addition to the capacity to act altruistically, some non-human primates have the capacity to adjust their behavior to whatever norms are in place as they move from community to community.[2] These ethologists rightly celebrate this second capacity as central to morality. It is worth noting, though, that humans have a further capacity that is no less central to morality: the capacity not merely to conform to norms (intentionally or not, from ulterior motives or not), but the capacity to do so because they judge conforming to be right. To discover of a group of primates (human or otherwise) that they successfully adjust their behavior to the norms that are in place is not yet to have evidence that they have this further capacity.

Kant famously highlighted this extra capacity when he emphasized the importance of acting from duty (as contrasted with acting merely in conformity with duty). Now Kant, of course, thought it made all the difference to the moral status of an action, holding that only actions done from duty have moral worth. I am not here signing on to that strong claim. But I do think the ability, not merely to do what is right, but to do it *because* it is right, is central to our conception of what a moral agent is. Moreover, I think the capacity, generally, to be guided by our judgments of what is right, or good, or justified is a distinctive capacity that we exercise not simply in acting morally but much more broadly in navigating through the world. Our capacity to be guided by our normative judgments – by our non-moral ones no less than our moral ones – figures crucially, I believe, in a proper understanding of our practical and our theoretical reasoning broadly construed. It is this general – not specifically moral – capacity that is my concern in this paper.

Incidentally, to hold that understanding this capacity is crucial to understanding *our* reasoning, is not to hold that the capacity is necessary to practical and theoretical reasoning *per se*. *We* regularly rely in our reasoning on judgments that deploy normative concepts and this makes our reasoning both more sophisticated and more self-reflective. But it is a contingent fact about us that we have these concepts at all and a contingent fact that our reasoning makes use of them. That said, as contingent as it is, we cannot even begin to evaluate our reasoning without normative concepts, so any attempt to understands ourselves critically will inevitably bring the concepts into play.

Agents who have this capacity to act on the basis of their normative judgments were valorized by Kant as "Rational Agents." As he recognized, an agent might have the relevant capacities and yet fail to act as her judgments would direct (and so fail to act rationally). Having standing as a "Rational Agent," in Kant's sense, is a matter of having the capacity to

[2] See, for instance, Frans De Waal's Good Natured (Harvard University Press, 1996). He has found nice evidence both that communities of non-human primates are able to introduce and enforce various social norms that seem to shape behavior by (at least) altering incentive structures and that individuals moving from one community to another can adapt quickly and well to differences in the norms. See also "Any Animal Whatever: Darwinian Building Blocks of Morality in Monkeys and Apes," by Jessica Flack and Frans de Waal, in Evolutionary Origins of Morality, ed. by Leonard Katz (Imprint Academic, 2000).

be guided by one's normative judgments, and one qualifies whether or not one exercises the capacity.

Kant goes on to offer distinctive and important views about what is involved in having the capacity, both when it comes to specifying what it is genuinely to make a normative judgment and when it comes to being able appropriately to act on those judgments. Specifically, but pressing things a bit out of Kant's own framework, (i) Kant thinks all genuinely normative judgments implicate universal principles that find their expression (in imperfect wills) in the Hypothetical and Categorical Imperatives and (ii) he maintains that rational agency requires an autonomy that consists not simply in successfully guiding one's behavior by one's normative judgments but also in being free of determination by alien causes.

In what follows, I will leave aside these views of Kant's, even as I take up the question: What has to be true of an agent for it to have the ability to act as it does because it thinks doing so is right, or good, or justified? What does it take, in other words, to qualify as what Kant calls a "Rational Agent"?

While I believe Kant was right that the rational agency is crucially important, I hope to spell out what rational agency requires in a way that steers clear of Kant's own appeal to hypothetical and categorical imperatives as well as his eventual reliance on noumenal selves and kingdom of ends. What follows is an attempt to articulate the idea of a "Rational Agent" with resources that are normatively and metaphysically more modest.[3]

Successive Approximations of Rational Agency

Kant introduces his view of rational agency this way: "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..."[4] He immediately goes on to treat the required conception of a law as the conception of something as "practically necessary, i.e., as good." In perfectly rational agents, he says, such representations are sufficient for determining the will. In less than perfectly rational agents, in contrast, the representations sometimes fail to determine the will and, because of this, the representations present themselves as commands or imperatives with which one 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 the undifferentiated "everything in nature" (that Kant mentions) 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 Kant specifies. Second, in the process, I hope to make clear that agents that lack the capacity to act as they do because they judge doing so to be good, or right, or justified, might nonetheless satisfy a variety familiar accounts of rationality. This tells neither against those accounts, nor against the Kantian account, but it does highlight well, I hope, the distinctive features of the specific capacity that Kant highlighted. 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 (what I am calling) rational agency to come on the scene.[5]

---

[3] At the same time, I hope the account of rational agency is compatible with Kant's more substantial commitments, even as it avoids taking them on.

[4] Grounding for the Metaphysics of Morals, Immanuel Kant (Hackett Publishing, 1993), translated by James Ellington, p. 23.

[5] On the side, so to speak, I hope that the characterization of rational agents that emerges makes it plausible that the metaphysics they would require is naturalistically tractable.

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 (but by no means all) of them represent the world. Thus photos, paintings, reports in newspapers, signs by the road, ideas, and animals, including humans, represent the world as being a certain way.

Needless to say, a lot is required in order for something to qualify as representing the world as being a certain way. And there is room, it seems, to contrast the ways in which photos, paintings, reports and signs represent the world, on the one hand, and the way animals, including humans, do, on the other. Perhaps the former qualify as representing at all only because of how they are used by the latter, while the latter might not themselves depend on anything else in order to qualify as representing what they represent. If so, then we could distinguish among the things that represent, those that do so only dependently and those that do so independently. In any case, I won't here explore what exactly is required in order for something to represent the world as being a certain way. Instead, I will simply assume there is some acceptable account and will proceed on the assumption that whatever that account is, it will allow the distinctions I will be relying on in what follows.

Whatever account of representation turns out to be right, something can represent things as being a certain way only if, also, it might misrepresent them. The risk of misrepresentation comes inevitably with the ability to represent. Talk of misrepresentation straightaway introduces a normative characterization. After all, any judgment that some thing is representing the world as being a certain way will rely on seeing that thing as subject to a norm of some sort (in light of which the representation might count as a misrepresentation). And this might raise a worry. If the things we are focusing on represent, and if representation comes with the possibility of misrepresentation, haven't normative concepts already come on the scene right at the start? In a sense, of course, they have. Our ability to think of something as representing (and so potentially misrepresenting) the world does require that we have a normative concept. Yet it does not require that the thing doing the representing have such a concept. While seeing something as representing involves seeing it as subject to norms, it does not necessarily involve seeing it as having normative representations.

In any case, whatever it takes to have the capacity to represent the world, a good variety of things seem to have whatever it takes. Among the things that represent, some act on the basis of the representations they have, moving or not as a result of how they represent the world as being.[6] We can, for instance, easily imagine building a little robot (out of Lego, say) that has light sensors and infrared transmitters the give it the capacity to represent various features of the world and that has the capacity too to respond differentially depending upon how it takes the world to be.[7] To the extent it is reasonable to see the robot as

---

[6] Something would presumably count as having the capacity to represent the world only if the putative representations changed in appropriate ways in response to information from the world as well as to other putative representations. And it may be that the changes in question are properly seen as themselves involving the agent acting in some way, e.g. inferring certain things from others. In that case, something would count as genuinely representing the world only if, at the same time, it counted as having the capacity to act (e.g. infer or conclude) on the basis of those representations. In that case it is an understatement to say that *some* things that represent act on the basis of their representations, since (on this view) all things that represent act on the basis of their representations. Of course, one might see whatever is required for representation as not itself requiring the capacity genuinely to act on their basis. Even then, at least some things that do represent the world also act on the basis of their representations.

[7] According to Lego advertisements, this is not merely something we can imagine. With the Lego Mindstorms sets, they claim "you can create everything from a light-sensitive intruder alarm to a robotic rover that can follow a trail, move around obstacles, and even duck into dark corners."

representing the light signals, obstacles, etc. as being on or off, or in the way or not, it will be reasonable to see the robot as having the capacity to act on the basis of its representations. Robots aside, all sorts of animals evidently have the capacity to respond to their representations of how the world is, relying on their representations to secure food, avoid predators, find mates, drive cars. Human and non-human animals alike depend upon this capacity in a variety of ways, some mobilizing very simple representations, others marshalling amazingly complex representational systems. What they share is the capacity to act on the basis of their representations.

For our purposes it is worth distinguishing among the various agents that have this capacity, marking, in the process, the ways in which the sophistication of the representations that govern their behavior matter to how we think of them. One thing that will emerge is that an agent might be tremendously sophisticated in thinking about the world and yet still lack the distinctive capacity to think of things as better or worse, right or wrong, justified or not.

Some agents that act on the basis of their representations are *stimulus-response agents*. They represent the world as being a certain way and then respond directly, without plans or strategies or any representation of alternative courses of action. They move left or right, stop and go, flea or freeze, etc. in response to their representations of the world being a certain way. Something could count as a stimulus response agent and yet have no representation of alternative courses of action, no representation of the response as achieving anything, and certainly no representation of the response as better or worse, as right or wrong, as justified or not. Something might count as a stimulus response agent while being cognitively very simple and while utterly lacking the capacity to represent things in normative terms.

That said, some stimulus response agents are more than merely stimulus response agents. Some agents have the capacity not simply to represent things as being a certain way, but also the capacity to represent the situation as being such that, *as a result of their own intervention*, things will turn out one way rather than another. Such beings can, in effect, represent different possible courses of action as being available and they have the capacity as well to respond differentially to those representations. On seeing that things are now a certain way, these agents – *planning agents*, I will call them -- see how they might be and respond differentially to the prospects, taking the course of action that attracts them the most or repels them the least. Needless to say, a tremendous amount of cognitive sophistication is on board before agents would qualify as planning agents in this sense. They need to be able to see themselves as facing different possible outcomes and they need too to see their own behavior as making a difference to what happens. With these resources available -- as long as they are combined with the capacity to act on the basis of these representations -- we have agents who can make plans, follow through on intentions, and maximize their expected utility. Planning agents, in fact, have all that it takes to satisfy the principles of standard decision theory. To the extent one identifies being rational with satisfying these principles, one will hold that a successful planning agent acts rationally. Yet a planning agent can be successful, and can consistently take those options it faces that maximize expected utility, without having the capacity to represent the various options as better or worse, right or wrong, justified or not. And if it lacks that capacity, then no matter how well it meets certain standards, it will not be acting as it does because it thinks acting in the way it does is good, or right, or justified. It will not yet qualify as a "rational agent" of the sort we are seeking.

Some agents, it is worth noting, are more than merely planning agents. Some agents have the 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 the ability to respond differentially to such complex representations, are *strategic agents*. They represent not just how things might be as a result of their own intervention, but also represent other agents as responding to their own representations of how still others will act in various situations. And they 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 strategic agency comes new possibilities and new problems. It introduces lying, for instance, as an option, since lying involves trying to get others to represent things as being a way in which one thinks they are not and this requires seeing others as representing the world and (presumably) responding to those representations. With the appearance of strategic agents comes the possibility of interactions among agents to which game theory applies. And strategic agents might well satisfy the principles of game theory, sometimes cooperating, sometimes not, as appropriate. Yet, again, such an agent can be successful, and can consistently satisfy the standards that emerge from game theory, without having the capacity to represent its various options as better or worse, right or wrong, justified or not. And if it lacks that capacity, then no matter how well it meets certain standards, it will not be acting as it does because it thinks acting in the way it does is good, or right, or justified. Even a strategic agent that satisfies the standards of game theory will not yet qualify as a "rational agent" of the sort Kant identified.[8]

Before moving on to *rational agents*, it is worth noting just how sophisticated the Strategic Agents might be without having the capacity to think in normative terms. 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. Or they might see themselves and others as having preferences and as responding to their options as they do in light of their preferences. Depending on how they think of preferences they might see themselves and others as inevitably choosing to do what they most prefer, or they might hold that sometimes agents take options that are not among those they prefer.

Overlaying these possibilities, strategic agents might well introduce rules for behavior with which they are disposed to conform and disposed to enforce in various ways. A community of such beings will have added something important to their conceptual repertoire, and in light of this acquisition they will have moved beyond being mere strategic agents to being, as we might say, *norm governed agents*. All of this is possible in the absence of a capacity to represent compliance *as good* and violations *as wrong*. It is one thing to represent a rule or principle, even where one is moved by that representation, and another to think of complying or failing to comply with that norm as good or bad, right or wrong.

---

[8] With these agents on board we have, in effect, all with which decision and game theory concern themselves (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 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.

Rational Agents

What more is needed, then, in order for an agent as sophisticated as strategic and norm governed 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 as well as the capacity to represent various norms as being in place 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 right or wrong, or as justified or not and (ii') be able to act on the basis of such normative representations. The crucial addition, of course, is the capacity to represent various options as better or worse, etc. Once that capacity is in place, the capacity to act on the basis of that representation is not significantly different in kind from the capacity to act on the basis of other representations, which is a capacity that has been on the scene from the start with stimulus response agents.

So what would have to be true of an agent for us properly to credit it with thinking of options as better and worse, etc. What would count as evidence that it possesses the relevant cognitive resources? Suppose that we were to come upon some community of primates (human or otherwise) that clearly count as at least strategic and norm governed agents, in the sense described above. What would count as evidence that they think of their options as better and worse (and are not simply more attracted to some than others)?

Deploying Our Normative Concepts

Cognitivists about normative thought see this question as, in important respects, nicely parallel to asking what would count as evidence that they think of some things in their environment as being blue and others as not? In both cases, the question is whether they have a certain sort of concept (normative concepts in the first case, color concepts in the second) and whether, assuming they do, they have some specific concept (of value in the first case, of blueness in the second).

Normative concepts certainly differ in kind from color concepts and specific normative concepts differ from one another just as different color concepts do. Yet, when the challenge is to discover if the thoughts of others might properly be characterized using the terms we would use to express our thoughts about value, these differences seem not to make an important difference to interpretative strategy. The thing to do (it seems) is, first, to deploy our own concept of value (or of blueness) to determine what, in their environment, is of value (or blue), and then, second, see whether they are sensitive to instances of value (or color) in a way that would underwrite attributing to them representations of things as being valuable (or as being blue).

Significantly, their being sensitive to instances of value (or blueness) in the relevant way is not the same as their reliably tracking value. On the one hand, an agent may reliably track value without having a concept of value at all (say, thanks to non-conceptual mechanisms). On the other hand, an agent may well have a concept of value and yet regularly, even systematically, get wrong what is of value. 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, given its situation, would be evidence of value. 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. In this case, as in the case of determining whether the agent has the concept of value, 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. Across the board, in fact, when we have grounds for attributing to someone some particular concept it is always because we have grounds for thinking that their representations are appropriately sensitive to the evidence they have that the concept in question is satisfied.

Of course, talk of being appropriately sensitive to evidence is extraordinarily vague, not least because whether some consideration or experience counts as evidence or not is itself 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. What is, for one agent, evidence that something is blue might well be for another, given different background beliefs and experiences, not evidence at all. And the whole process is complicated even more because it often becomes plausible to attribute some particular concept (say, of blueness) to some agent only as it becomes plausible too to attribute a range of other concepts.

Nonetheless, our own competence with the concepts in question, combined with an appreciation of the situation of the agents in question, put us in a position sometimes to make judgments about what evidence is available to them and so, also, about whether they are appropriately sensitive to the evidence they have that the concept is satisfied. Of course, we are not always in a good position to make such judgments. And even when we are reasonably well placed, there is plenty of room for mistakes. But, in any particular case, the justification we have for attributing the relevant concepts to others will rise and fall in tandem with our having reason to see them as appropriately sensitive to the evidence they have.

It is worth noting that while our grounds for attributing concepts to others depends on our having reason to see them as appropriately sensitive to the evidence they have – and so requires our having the (apparently normative) concept of evidence – the agents we are seeking to interpret need not have any such concept. Being appropriately sensitive to evidence that a concept is satisfied does not require having, among one's concepts, the concept of evidence. So while the concept of a concept is, at least as I am approaching things, bound up with the concept of evidence, there is (so far) no reason to think one can have concepts only if one has the concept of evidence.[9]

The underlying idea 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 appropriately sensitive to evidence (to the reasons there are for thinking the concept in question applies). Since the characterization of the functional state is in normative terms, there is no promise here of reducing the normative to the nonnormative. All the same, it is worth noting that the dispositions that would allow the characterization to fit appear not to involve any mysterious metaphysics or occult sensitivities. Nor do those dispositions necessarily involve the capacity to represent the evidence in normatively loaded terms as evidence.

A Crucial Contrast Lost

The very fact that, so far, our grounds for attributing a normative concept to agents are so similar to the grounds for attributing a non-normative concept to them raises an important worry: that this approach cannot capture what is distinctive about our normative concepts of value, rightness, and justification.

---

[9] Moreover, 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.

Here is a way to press the worry. Suppose one were inclined to hold, as many have, that being valuable is a matter of being such as to secure approval from someone under certain circumstances (e.g. from someone who is fully informed and impartial).[10] With such an account in hand, the challenge of determining whether some agents have the concept of value becomes the problem of determining whether they are appropriately sensitive to evidence that things would secure the relevant approval under certain circumstances. As I noted above, this does not require that their representations reliably track what would secure that approval. Rather, what needs to be true of them is that their representations are deployed in response to the evidence they have (which may well be misleading) that things would secure that approval. With this account of value in mind, imagine that we discover of some agents that they are, in fact, sensitive in the appropriate way to evidence that things would secure approval under the relevant conditions. Have we then good grounds for attributing to them the concept of value? Well, it is at least tempting to think not. Despite all their appropriate sensitivity to evidence, it seems that the agents might still have only a non-normative concept of a disposition – the disposition to secure approval – and not a concept of value at all. Put another way, while they have a concept that is a concept *of what is valuable* (if the account of value is right), they may well not have a concept of it *as valuable*. They will count as having the latter, it seems, only if the concept they are deploying is a normative concept.

Deploying Our Concept of a Normative Concept

What does it take for a concept to be a normative concept? This question suggests the second approach one might take to determining whether some agents have the capacity to represent things as valuable. 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 a normative concept. Against the background of having established that an agent is appropriately sensitive to evidence concerning which things are valuable, the question is whether the concept there in play (that is differentially applied in response to the available evidence) is a normative concept or not. The guiding thought is that the agent is properly credited with thinking of the valuable things *as valuable* only if concept is a normative concept. And to determine whether it is, we need to deploy our concept of a normative concept to see whether the agent's concept qualifies as one.

With this in mind, it is important to identify, if possible, what is distinctive of normative concepts. In virtue of what does a concept count as a normative concept? Unfortunately, it is difficult to say just what our concept of normative concepts requires.

A common suggestion, though one that seems inadequate, is that a concept counts as normative if, and only if, it is action guiding in some appropriate way. Usually, the idea is that a concept counts as action guiding in the relevant way by having a (not necessarily decisive) impact on action. On this view, to see something as (for instance) good, is, in effect, to be attracted by one's representation of it, where the content of that representation can be spelled out in nonnormative terms. Thus, for example, to be someone who thinks that honesty is good is to be someone who is motivated appropriately by the thought that some course of action is honest or dishonest. And the difference between a person who sees honesty as good and one who does not is found, on this account, in the different motivational role that representations of honesty play in those people.

---

[10]. David Hume (in the Treatise of Human Nature) and Adam Smith (in The Theory of Moral Sentiments) developed views of this sort. For more contemporary versions see Roderick Firth's "Ethical Absolutism and the Ideal Observer" in *Philosophy and Phenomenological Research*, vol. 12, No. 3 (1952), pp. 317-45 and Michael Smith's The Moral Problem (Blackwell, 1994).

The inadequacy of the suggestion comes out clearly with reflection on the various representing agents – the stimulus response, planning, strategic, and norm-governed, agents – that fall short of being Rational Agents (i.e. fall short of doing what they do because they represent it as good or right or practically necessary). These agents 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 normative concepts, nor are the agents, simply in virtue of the motivational impact of their representations, properly credited with thinking of things as better or worse, etc. Thus, for instance, when a simple stimulus response agent develops the disposition to avoid red things, the concept of redness hasn't then become a normative concept, just a causally efficacious concept. Similarly for agents that are moved to take options that they represent as resulting in pleasure, or as conforming to norms that are in place. The representation of future pleasure, or of conformity with a norm, are in such cases causally effective, but that is compatible with the agents utterly lacking the capacity even to think of the pleasure or the conformity as good or bad, right or wrong, despite the impact the representation of pleasure or conformity might have on their behavior. Also, of course, there are various arguments for thinking that an agent may in fact have normative concepts and yet be such that those concepts are not, for them, actually action guiding. These suggest that having some particular motivational role is neither necessary nor sufficient for a concept counting as a normative concept. An attractive alternative would be to hold that the status of a concept as a normative concept is not tied to its actual motivational role but to the motivational role it would have in those who are rational or responsive to reasons. In the end, I embrace something along these lines, but for right now the important point is that the motivational impact of some concept does not establish it as a normative concept, no matter how consistently it guides behavior. Something more, or different, is required.

What then is required for a concept to count as a normative concept (and so, as a concept that at least might be a concept of value or rightness or justification)?

I would like to back-in to a proposed answer by discussing first a common objection to dispositional accounts of value. According to this objection, such accounts will seem plausible only if the things that would 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.[11] Good things are such that they do not merely cause approval (from those appropriately situated), they merit the approval. Yet, the objection presses, that means various dispositional proposals will seem plausible only so long as one is implicitly relying on some independent criterion of value in light of which things are thought to merit the approval. And the need for an independent criterion belies the theory's claim to having accounted for (as opposed to presupposed) value.

This is too quick, though. Someone attracted by the dispositional theory can perfectly well grant that some specific version of the theory is plausible only if the things that would garner approval from those in the situation that version privileges merit the approval and yet also hold that the standard for whether they merit the approval they receive is the very same standard applied to itself. To ask whether those things that secure approval under the specified circumstances *should* secure that approval, is, according to such a view, to ask whether the fact that those things secure approval under those circumstances would itself

[11] This is a point John McDowell makes in "Values and Secondary Properties," in Morality and Objectivity, Ted Honderich, ed., (Routledge and Kegan Paul, 1985), pp. 110–29, when discussing the suggestion that dangerousness should be understood in terms of dispositions to prompt fear.

secure approval under those circumstances.[12] The problem with the objection is that it assumes, without grounds, that the dispositional theory would need to appeal to some independent criterion of value in order to determine whether various responses were merited.

As long as the original pattern of approval would itself secure approval under the specified circumstances, those original approvals would (as the dispositional theory would have it) count as good, or merited, or justified, under the circumstances. This means that, at least in principle, someone who embraces the dispositional account of value could argue consistently that the approval that is being taken as the standard of value is an approval that is itself certified (by that very standard) as good, or merited, or justified. And this means a dispositional theory can consistently acknowledge that something counts as valuable only if, in addition to securing approval under certain conditions, that approval is itself justified. There is no need to appeal to an independent standard to make sense of the idea that the approval is justified – the standard offered by the theory might play that role.

As long as there is reason for thinking the second order approval would be forthcoming, the dispositional account of value can accommodate the demand that the things that prompt the approval should merit that approval.

Admittedly, this response – that the approvals themselves secure the appropriate approval – has an aura of triviality. Yet there is nothing trivial here. Whether the approvals in question would themselves in fact secure approval is a substantive question. 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 a particular version of the dispositional account survives the test, it accomplishes no small feat. For what it is worth, it would not be surprising if fairly often, as people reflect on their own patterns of approval they discover aspects of themselves of which they don't approve, just as, when they reflect on what scares them, or excites them, or makes them uncomfortable, they often don't approve of their own reactions. Whether a certain sort of approval, garnered under certain special conditions, might itself secure that approval under those conditions, is an open question.

Right now, though, my interest is not in whether a particular dispositional account satisfies this test, but in the relevance of the test itself. Why think a particular dispositional account would be plausible only if the sort of approval it treats as defining value would itself be merited? What would be wrong with a dispositional theory that simply rejected as irrelevant the question of whether the specified approvals were merited?

I think that getting a good answer to this question reveals something deep and important about our normative concepts. Unfortunately, as convinced as I am of this, I have more than a little difficulty articulating a good answer. I will, nonetheless, do my best.

The first thing to do is to note what would be wrong with a dispositional theory that failed the test. In that case, the theory would be saying roughly, first, that certain things are in fact good (because they would garner approval under the specified circumstances) while also saying, of those very things, that there is nothing good about them being good – that, from the point of view of value, it would have been just as good had something else been valuable.

Here is a different way to describe the situation: the theory would be holding that there is no justification for (i.e. nothing valuable about) using the criteria of value it advances for distinguishing between what is valuable and what is not. What the dispositional theory is

---

[12] I am not trying here to defend this view as acceptable but only to show that it has the resources to meet, on its own terms, an objection that many treat as decisive.

doing is offering a particular standard as being such that satisfying it is both necessary and sufficient for counting as valuable. If that theory's own standard doesn't meet the standard on offer, then there is (on this theory's account) nothing valuable about meeting the standard. And if there is nothing valuable about meeting it, then the fact that something meets it does not after all show that there is anything valuable about the thing in question. But if meeting the (putative) standard of value does not *ipso facto* establish the value of what meets it, then the standard cannot be the right standard.

So it seems. What is going on here? Well, first of all, I believe the relevance of the test reveals a distinctive feature of normative concepts -- that the standards for their application are always in principle themselves open to evaluation -- and answerable to the results. If a concept is a normative 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 in deciding how to act and react.

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 the standard to be used in determining what is of value. And the answer we come to is probative with respect to whether we have the criterion right: a negative answer provides grounds for the standard is not actually the right standard.

There is an important contrast, here, with non-normative 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. For instance, once we have the concept of blueness up and running, to ask of the standard for its application whether we should rely on that standard in making choices 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 a normative 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.[13] To discover that the standard is one that implies distinctions we cannot justify as important is to discover it is not, actually, the right standard for determining what is (and is not) good.

Of course, there are a lot of concepts that are clearly not normative concepts that nonetheless are such that we can ask, of the criterion for their application, whether we have the criterion right. And the answer we come to will be probative with respect to whether we should accept or reject the criterion. So not just any evaluation of a criterion for application is relevant to revealing the normative nature of a concept. What sort of evaluation needs to be possible and probative, in order for a concept to count as normative?

To answer that, I think we should appeal to an initially not very informative, but for that reason not very controversial, observation concerning normative concepts: a normative concept is a concept the satisfaction of which provides reasons to someone or other to do

---

[13] It is worth noting that the test here on offer is not a reflexivity test. The question is not, of each normative concept, does it satisfy itself. When it comes to the concept of badness, for instance, which is just as much a normative concept as that of goodness, the crucial point is that challenges concerning the value of the standard we use in applying the concept of badness are probative with respect to our having the right standard.

something or other.[14] Put another way: normative concepts are such that when things (actions, options, objects, people) satisfy them, there is *ipso facto* reason for someone to do (or refrain from doing) something or to be (or avoid being) a certain way.[15]

A candidate criterion for some normative concept will be one that is offered as being such that meeting it means there is reason to do or refrain from doing something. So, for instance, if the concept in question is that of being approvable, a particular criterion on offer, to be successful, must be such that satisfying it provides reason to approve of whatever satisfies the criterion. Evidence that one would not be justified in approving when the criterion is satisfied is evidence that the criterion does not, after all, capture what it takes to be approvable. Or, to take another example, if the concept in question is that of being a duty, a particular criterion on offer, to be successful, must be such that satisfying it provides decisive reason for the person with the duty to act accordingly. Or to take still another example, if the concept in question is that of being disgusting, a particular criterion on offer, to be successful, must be such that satisfying it provides reason (not necessarily decisive, nor even strong) to feel disgust.

In each case, the test of a proposed criterion for the application of a concept is whether that criterion sustains the necessary connection between something satisfying the concept and reasons. The same test, with different sorts of doings or refraining, and different sorts of reasons, at stake, can (I think) be said about all normative concepts.

Thus to evaluate a proposed criterion (in the relevant way) is to ask whether something satisfying it is, in itself, reason for someone to do or refrain from doing something. If the concept in question is a normative concept, to discover that something satisfying the proposed criterion does not, in itself, mean there is reason for someone to do or refrain from doing something, is to have decisive grounds for rejecting the criterion as wrong. A proposed criterion for the application of a normative concept cannot be correct if satisfying that criterion does not, *ipso facto*, provide reason to someone or other to do something or other.

Normative Concepts

Normative concepts are such that their satisfaction *ipso facto* provides reason. This means that competence with a normative concept requires being sensitive to evidence that the criteria for its application is such that its satisfaction *ipso facto* provides reason. And being sensitive in this way requires that one's deployment of the concept reflects a willingness to adjust one's criteria for its application in light of evidence that the criterion upon which one was relying is not actually such that satisfying it *ipso facto* provides reasons.

Sometimes, this willingness is manifested in shifting one's normative view in light of arguments offered by others, other times it is shown in resisting those arguments with arguments of one's own. Taking seriously the idea that one's own understanding of the

---

[14] If all concepts are such that their satisfaction itself provides reason to believe they are satisfied, one might stipulate that the reasons provided by the satisfaction of normative concepts must be reasons to do something over and above merely believing the concept is satisfied. Alternatively, one might hold that our concept of satisfying a concept is itself a concept of something normative. In that case, a particular proposal as to what counts as 'satisfying a concept,' to be successful, would have to be such that satisfying it provides people with reasons to believe. I am not sure which way would be better, nor am I sure that concepts are such that their (mere) satisfaction itself provides reason to believe anything.

[15] Certainly, non-normative concepts are also such that, when things satisfy them, there is sometime reason to do (or refrain from doing) something. Yet when this is true the connection between satisfying the concept and there being a reason is contingent.   The fact that something is blue may well give me reason to choose it, but only in light of other considerations (some of which are normative).

normative concepts one uses is liable to challenge and can stand vindicated only with the support of arguments showing that the criteria one relies offer justifiable grounds for acting (or not) is a natural expression of competence with normative concepts.[16]

The liability to challenge goes hand in hand, I think, with the sort of claim to legitimacy that normative concepts carry. Thus to discover of some population's concept that it is not liable to such a challenge, that those possessing it would reject as out of place the question of whether what satisfies the standards in play are good, is (again) to find grounds for thinking the concept they are using is not a normative one

This fact about normative concepts helps explain two familiar features of normative concepts: (i) that they are, as W.B. Gallie put it, essentially contestable[17], and (ii) that for any normative concept one can always, as Moore pointed out[18], intelligibly challenge as wrong any proposed standard for its application. While neither of these features seems unique to normative concepts, they are both characteristic of normative concepts and a relevant challenge to any account of normative concepts is that it can make sense of why such concepts all have these features.

Both features of normative concepts reflect the way in which competence with a normative concept goes with being sensitive to the evidence that satisfying it *ipso facto* provides reason to do (or refrain from doing) something. Specifically, two clear ways of manifesting an insensitivity to this evidence is to treat the standards embodied in concept's deployment as uncontestable and to reject unintelligible the question of whether satisfying those standards actually provide reason. If someone does either we have grounds for doubting they possess the normative concept in question.

These grounds are not, of course, decisive. A person might properly be credited with a concept even it she fails to be sensitive to some of the relevant evidence that is available to her, so too with a group of people. Concept possession cannot plausibly require perfect sensitivity to relevant evidence; there are too many good explanations of why people who clearly possess some concept fail to respond appropriately to the evidence they have. Still, if people fail consistently to regard some evidence as relevant to a concept they are deploying that is reason to suspect their concept is not one to which that evidence is relevant. So, when it comes to normative concepts, if people consistently treat, as irrelevant, challenges to the justifiability of their standard for some concept, that is reason to suspect their concept is not a normative concept.

To put things in a slightly different way: just as our grounds for attributing various familiar non-normative concepts involve discovering whether those who supposedly possess it are appropriately sensitive to evidence for its applicability, so too with normative concepts. The crucial difference is that, in the case of normative concepts, the relevant evidence includes the justifiability of the standards in play.[19]

---

[16] While engaging in normative discussion is, as I say, a natural expression of competence with a normative concept, it is worth emphasizing that a person might be competent with a normative concept and yet, for various reasons or causes, resist such discussion. Moreover, one might exhibit the requisite sensitivity to evidence concerning reasons without having the sophistication required to engage in normative discussions. The discussions are one mark of the relevant competence, not a necessary condition for it.

[17] "Essentially Contested Concepts," *Proceedings of the Aristotelian Society*, Vol.56, (1956), pp.167-198.
[18] Principia Ethica (Cambridge University Press, 1903).
[19] Perhaps it is worth mentioning too that if I am right about the nature of normative concepts it is no surprise that the method of reflective equilibrium is well suited to the job of articulating and defending normative concepts. The method reflects directly the way in which, when it comes to normative concepts, relevant evidence as to whether it is being correctly applied is found in whether the application in question preserves

There is, I think, a third feature of normative concepts that is much less familiar (and also hard to pin down) but worth mentioning. To introduce it, let me start by distinguishing *a better theory of X* from *a theory of a better X*. In thinking about the law, for instance, we might be comparing theories of what the laws (say, of the United States) are, and defend one over the others as a better theory of what the law is. Alternatively, and this is quite different, we might be comparing various theories of what the law should be and defend one over the others as being a theory of a better system of laws. The distinction seems pretty clear. Moreover, thinking through which would be a theory of a better legal system is irrelevant to the question of which theory is a better theory of the system that is in fact in place. If I settled, tentatively, on some view of what the laws in our society are (say, concerning same-sex marriage, or the right to abortion, or whatever) and someone convinced me that things would be better were the law different that I take it to be, that would provide no pressure for me to revise my understanding of what the law is. It would only provide pressure for me to work to change the law.

More accurately, and significantly, it would provide no pressure for me to revise my view of the law *as long as* my conception of what the law is eschews an appeal to normative concepts.

I might, however, hold a view (a la Dworkin[20]) according to which, on the best theory of what the law is, it is a system of principles that are answerable to demands for equal concern and respect. In that case, the distinction between a better theory of the law and a theory of better laws is elided, precisely because (on this view) to hold a view on what the law is, is to take a normative stand. To discover that one law would be better (when it comes to considerations of equal concern and respect) than what I have taken the law to be, provides some grounds for shifting my view of what the law is.

Of course, it only provides some grounds, not necessarily sufficient grounds. Even if, as Dworkin argues, our legal system is infused with normative concepts that make the law answerable to moral arguments, there are aspects of the law that ensure that there will always be room to distinguish between what the law actually is and what it should be, and so between the best theory of what the law is and a the theory of what the best law would be.

In thinking about the principles of morality, though, one cannot sustain a distinction between a better theory of morality (where this is understood as a theory of what the principles of morality actually are, not a theory of what people believe them to be) and a theory of a better morality. If I settled, tentatively, on some view of what the principles of morality are (say as they concern same-sex marriage, or allowing abortion, or whatever) and someone convinced me that things would be better were the principles of morality different than I take them to be, that would immediately put pressure on me to revise my understanding of what the principles of morality are.

The same is true of theories of justice, theories of virtue, and theories of rationality. In each case -- and in contrast with theories that do not rely on normative concepts – figuring out which theory is a better theory of the area in question is sensitive to whether things would be better were they different than the theory supposes them to be. To discover things would be better were they otherwise than the theory supposes is to find grounds for rejecting the theory in favor of an alternative.

---

the necessary link to reasons for someone or other to do something or other. If the link is not preserved, if the particular standard of application on offer is not such that something satisfying it *ipso facto* provides reason, then that is decisive evidence against the standard.

[20] Taking Rights Seriously (Harvard University Press, 1977).

This all reflects the fact, I think, that a neat and important mark of normative concepts is that, when they are in play, the distinction between a better theory of X and a theory of a better X simply is at least blurred and, when only normative matters are at stake, utterly collapses. That this is so is not surprising if I am right that a concept counts as a normative concept only if both (i) the criteria used in its deployment is open to evaluation and (ii) that evaluation is probative with respect to whether the criteria used are correct.[21]

Conclusion

Imagine that we were to come upon a community that used a term that sounded a lot like "right," that was applied regularly to actions that were, as we see them, actually right. Imagine too that they do this on the grounds that those actions met some social norm that was in place, a norm that we see as appropriate for their circumstances. Thus they not only judge to be "right" actions that we think are right, they also, in appealing to the relevant norm, articulate considerations in favor of thinking it 'right' that coincide with the considerations we would marshal for thinking it right. We would reasonably think that they are deploying the same concept we are when we judge of things that they are right. After all, they are, with their concept, picking out what is in fact right.

Yet imagine too that 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. We might see, for instance, that as the norms shift their use of the term shifts accordingly and without any thought to whether the new norms are better or worse than the old ones upon which they had been relying. If we did discover this, we would have grounds for suspecting that their concept is not actually a concept of rightness, and indeed not a normative 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 are compatible with or even suggest 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 normative concepts necessarily welcome or respond appropriately to demands for justification. Regularly they do not. Yet if the concept in play is genuinely normative 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 agents that are merely strategic or norm-governed, on the one hand, and Rational Agents, on the other, my suggestion is as follows. For agents to be Rational Agents they must possess and be appropriately responsive to normative 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

---

[21] Two other familiar phenomena bear mention as fitting well with, and perhaps being partially explained by, seeing normative concepts in the way I am suggesting: one is the ever openness of Moore's open question, the other is the evidentiary force of the method of reflective equilibrium. I will not here, now, go in to either.

responsive requires not reliable tracking but responsiveness to evidence of the relevant sort. So much we get as a requirement by applying a general account of what has to be true of agents in order for them to be properly credited with any sort of concept (normative and non-normative concepts alike).

When the relevant dispositions are in place we have grounds for thinking a concept is in play, we can render intelligible a distinction between how things seem and how they are, and so can see the agents in question as 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 a normative concept. And that means, if I am right about our concept of normative concepts, that their deployment of the concept in question must be sensitive not just to whether things satisfy a certain standard but also whether what satisfies that standard is such that, *ipso facto*, there is reason for someone to do or refrain from doing something or other.

Perhaps it is worth emphasizing that such agents need not themselves have a concept of a reason, nor do they need to be engaging in a meta-level reflection on their own concepts. What they do need to be doing is adjusting their own standards for deployment of their concepts in a way that is responsive to whether they have evidence that satisfying the standard they are relying on is, *ipso facto*, reason providing. So while this account of normative concepts makes essential appeal to the concept of a reason, and treats appropriate responsiveness to (evidence concerning) reasons as a defining characteristic of normative concepts, agents may well have normative concepts without having the concept of a reason. Moreover, while the account of normative concepts relies heavily on the idea that those who have normative concepts have concepts the standards of application of which must be sensitive to evidence that relying on them would not be justified, there's nothing in the account that implies that one could have a normative concept on if one could either in thought or discussion actually justify the standards upon which one relies. It must be that one relies one the standards one does in a way that is responsive to evidence one has concerning whether satisfying those standards provides reason, but one can do this without having the cognitive resources that would be required to work out or offer justifications for those standards.

That said, when one is interacting with someone who does have the capacity to reflect on and articulate a justification of the standards she relies on in deploying a concept, if we discover that whether she thinks she has the right standard of application for that concept is insensitive to questions of what agents have reason to do, we have grounds for suspecting the concept she is using is not a normative concept.[22]

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 normative judgments of their options, they qualify as Rational Agents in the robust sense that they act as they do because of their judgments of what is, and is not, worth doing.

---

[22] Think of the sort of evidence that would lead one to see someone as judging things as 'good' (in what is called the inverted commas sense) and not as good.