

Rational Agency and Normative Concepts

by

Geoffrey Sayre-McCord

(UNC/Chapel Hill)

Introduction

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. More generally, there is a difference between conforming to norms (intentionally or not, from ulterior motives or not) and doing what one does because one judges it to be morally good or right. The difference is, I think, central to morality and my main interest here is to get a handle on what has to be true of people for them to do what they do because they think it right or good. The abilities required are, I think, a special case of the abilities that are required to be what Kant identified as a rational agent. I focus on this more general capacity (the having of which is a necessary condition of moral agency) in the rest of the paper. As will become clear, I think Kant was right that the rational agency is important. I hope, though, 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 underwrite Kantian convictions (concerning rational agency) with more or less Humean resources.

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..."¹ 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

¹ Grounding for the Metaphysics of Morals, Immanuel Kant (Hackett Publishing, 1993), translated by James Ellington, p. 23.

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.² 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 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.

² 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.

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*.³

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

³ 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.

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, 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 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-normative 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 a normative 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 a normative 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 a normative 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 a normative concept.

Fair enough, I think. But at this point 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 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 normative concepts. For instance, when an 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. The representation of future pleasure is in this case causally effective, but that doesn't mean that it is a normative 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-normative concept of a property with a normative 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 must be such that they *should* secure that approval, that such a response is *good* or *appropriate* or *justified* under the circumstances.⁴ 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. As long as a person who embraces the dispositional account has independent reason for thinking it plausible, she has the resources to respond to

⁴ 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 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 triviality. Yet it is not. Whether the approvals in question would 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 wouldn't surprise me to discover that 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, it seems to me, 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. Why think a particular dispositional account would be plausible only if the sort of approval it treats as defining value would itself secure that approval? 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, 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, perhaps better, had something else been valuable.

Here is a different way to describe the situations: 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 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 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 I think. What is going on here? Well, first of all, I believe the relevance of the test reveals a distinctive (and largely unrecognized) 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.

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.

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.

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 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.⁵

Of course, there are a lot of concepts 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 accept or reject the criterion. So not just any sort of probative evaluation 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: that

⁵ 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 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.

normative concepts are such that when things (actions, options, objects, people) satisfy them, there reason to do (or refrain from doing) something -- where what one has reason to do would have to be something other than just believe the concept is satisfied. 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 on the grounds that the criterion is satisfied is evidence that the criterion does not, after all, capture what it takes to be approvable. Similar things, with different sorts of doings or refraining at stake, can (I think) be said about all normative concepts. To evaluate a proposed criterion (in the relevant way) is to ask whether something satisfying it is, in itself, reason to do or refrain from doing something. And to discover it is not is probative with respect the claim that the criterion is correct.

This feature of normative concepts both reflects and in a sense explains two features of normative concepts that are worth mentioning. The first is, I think, quite familiar, the second not. Both, though, work to highlight how the distinctive way in which normative concepts are liable to challenge plays out.

The first feature is that normative concepts are essentially contestable. 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 normative 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-normative concepts involve discovering whether those who supposedly possess it are appropriately sensitive to evidence for its applicability, so too with normative concepts, though in the case of the normative concepts the relevant evidence concerns the justifiability of the standards in play.

The second feature, as I say, is less familiar. 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 the law and defend one over the others as a better theory of what the law is. Alternatively, though, we might be comparing various theory 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) 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. 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, is grounds for shifting my view of what the law is.

Similarly, in thinking about morality, one cannot sustain a distinction between a better theory of morality 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 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 they are. 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.

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 cannot be sustained. That this is so is not surprising if I am right that a concept counts as a normative concept only if both the criteria used in its deployment is open to evaluation and that evaluation is probative with respect to whether the criteria used are correct.

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 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 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. If we did, we would have some grounds for suspecting that their concept is not 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 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 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 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 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 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 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 that standard is itself justified (which is to say, at least roughly, that the standard itself is such that it satisfies the standard of value in play - and its use as a standard is 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...