THREE KINDS OF MORAL RATIONALISM[*]

Michael Smith

1. Background

Moral rationalism can be understood in many different ways, but here it will be understood as the view that moral facts entail facts about moral reasons for action—more on what *moral* reasons are presently. Since there are many different ways in which moral facts could stand in entailment relations to facts about moral reasons, and since some of these are more plausible than others, let me begin by saying a little about what these entailment relations might be (here I find myself largely in agreement with Portmore 2011).

It seems to me that the most plausible form of moral rationalism will hold that an act's *being morally forbidden* entails that there are moral reasons, that some of these are reasons not to perform the action, and that these moral reasons are weightier than the moral or non-moral reasons to perform the action; it will hold that an act's *being morally permissible* entails that there are moral reasons, and that it isn't the case some of these are reasons not to perform the action that are weightier than the reasons, whether moral or non-moral, to perform it; and it will hold that an act's *being morally obligatory* is a matter of its being uniquely morally permissible. So understood, the connection between these deontic statuses of actions—being morally forbidden, morally permissible, and morally obligatory—and the existence of moral reasons to perform or not perform those actions is somewhat indirect.

For example, there might be moral reasons not to perform a certain action, but because those moral reasons are not weightier than the non-moral reasons that there are to perform it, the action turns out to be morally permissible. Imagine that while on my way to catch a train to see a play I've always wanted to see, I come across someone struggling with some parcels who needs help, but providing the help would make me miss my train and so prevent me from seeing the play. If the reason to provide the help does not outweigh the reason to see the play, then on the proposed account of the entailment relations, my not providing help will turn out to be morally permissible in these circumstances, notwithstanding the fact that there is still a moral reason to provide that help. The important point about morally permissible actions, on this way of understanding moral permissibility, is that morally permissible actions are those that moral reasons don't tell against *decisively*, not those that there are no moral reasons against performing those actions at all.

Moreover there might be two different actions that moral reasons don't tell decisively against performing where the performance of one is supported by non-moral reasons and the other by moral reasons, so both are morally permissible, but the weight of these reasons is such that one of these has a further deontic status beyond that of being merely morally permissible. To return to the example of the parcels and the play, imagine that the weight of the reason to see the play is perfectly balanced with the weight of the reason to provide help, or greater in weight. Providing help and seeing the play would in this case both be morally permissible, as moral reasons don't

[*] Many thanks to Karen Jones and Francois Schroeter for their very helpful comments on the first draft of this paper. Thanks also to Tristram McPherson for the many conversations we have had about this topic over the years, and to the participants in a seminar I co-taught with Thomas Schmidt for graduate students from the Humboldt University and Princeton University in the summer of 2016.

tell decisively against either helping or seeing the play, but since it would be morally better to perform the action that is supported by the moral reason, providing the help would have a deontic status beyond that of being merely morally permissible, namely, that of *being supererogatory*. Certain morally permissible actions, those that are supererogatory like providing help to the person struggling with their parcels in the circumstances just described, may add to the agent's moral credit, whereas other actions available to the agent that are similarly morally permissible like seeing the play do not. Indeed, since a supererogatory action might add to an agent's moral credit even when acting on the relevant moral reason isn't what the agent has all things considered reason to do, as he has all things considered reason to act on a non-moral reason instead, an agent's performing a supererogatory action could be a silly or perverse thing for an agent to do (compare Susan Wolf 1982 on moral saints).

To fully spell out a form of moral rationalism, two further questions would need to be answered. The first is what the difference is between moral and non-moral reasons for action. As is perhaps clear from my examples, this terminology is somewhat unfortunate, as though the difference lies in the nature of the reason-giving features, in both cases these features may be non-moral in nature, in the sense that they can be characterized without using any moral concepts. What makes moral reasons for action *moral* reasons is rather that, however these features need to be characterized in order to bring out their reason-giving nature, they must be impartial—in other words, they must be features of agents as such, not just features of the agent himself —and they must provide reasons unconditionally, not conditionally on the fact that the agent has certain desires, or some other non-rational disposition to acquire desires. In the case of non-moral reasons for action, these two conditions need not be met by reason-giving features. (It should now be clear why I characterized the fact that an act would be a helping of someone who is struggling with parcels as a moral reason to perform that action, and the fact that an act would enable me see a play I've always wanted to see as a non-moral reason to perform it.)

The second question that would need to be answered, in order to fully spell out a form of moral rationalism, is the modal and epistemic status of the moral rationalist's claim. Moral rationalists hold that whenever moral facts obtain, certain other facts obtain, facts about reasons for action that are impartial and unconditional. But they could hold that this is a contingent a posteriori truth, or a contingent a priori truth, or a necessary a posteriori truth, or a necessary a priori truth. In what follows the focus will be on those forms of moral rationalism that hold it to be a necessary and a priori truth. As I understand it, this is the modal status of the moral rationalist's claim on all standard versions of the doctrine. Non-standard versions of moral rationalism according to which the connection is contingent and a posteriori, or contingent but a priori, or necessary but a posteriori, will be ignored in what follows.

So much for scene-setting. Suppose we grant the moral rationalist that it is necessary and a priori that wherever moral facts obtain facts about reasons for action that are impartial and unconditional obtain. It is then fair to ask them why this is so, and what its being so tells us about morality more generally. My task in what follows is to talk through three very different answers moral rationalists could give to these further questions. The three answers correspond to three very different formulations of moral rationalism, and these different formulations in turn have surprising implications for whether moral knowledge is vulnerable to an external challenge of the kind mounted by John Mackie (1977) and Richard Joyce (2002). As we will see, two of these formulations, the two that are subject to decisive objections, make an external challenge impossible, while the third and best formulation leaves that possibility wide open.

The aim of the paper is thus three-fold. The first aim is to put forward a general characterization of moral rationalism in terms of moral facts entailing facts about moral reasons for action, and then three more precise formulations of the doctrine that are supposed to help us see why this is true; the second is to argue that one of these formulations is superior to the others; and the third is to explain how the arguments of those moral rationalists who claim that moral knowledge is invulnerable to an external challenge depend on their preferred inadequate formulations of moral rationalism (see especially Dworkin 1996; Scanlon 2014—but for a different kind of response to these moral rationalists see McPherson 2008, 2011).

2. Epistemology for moral rationalists

The canonical way in which we find out what the moral facts are—the way given early expression by Aristotle in his quite general account of philosophical method, developed in detail for the moral case by John Rawls (1951, 1971, 1974), and now widely employed and taught in normative ethics courses—is by engaging in reflective equilibrium reasoning. We fix on specific cases in which we are confident of what the moral facts are and the general moral principles that we think likely justify our judgements about those specific cases, and we then test each of these by extending the general principles to other specific cases and seeing whether the judgements that they lead us to make about them are similarly credible.

If they are—that is, if our judgements about general principles and specific cases are in reflective equilibrium with each other, and there is no pressure to revise either in the light of the other—then, if these judgements are also in wide reflective equilibrium with the rest of our beliefs about the world, we *may* already have moral knowledge—more on that 'may' presently. But if they aren't, then we must adjust either our initial judgements about specific cases, or our general moral principles, or some of the rest our beliefs about the world, in an attempt to bring all of these into a wide reflective equilibrium with each other.

According to moral rationalists, moral truths are knowable a priori, so the judgements about specific cases that we bring into reflective equilibrium with the general principles that we think likely justify them, and those general principles too, are immune to empirical refutation. They should therefore be thought of as hypothetical in form—'If the non-moral facts are thus-and-such, then the moral facts are such-and-so'—rather than as being in part empirical speculations about the non-moral features possessed by particular actions performed by particular people at particular points in space and time.

Moreover, since we might originally have acquired such beliefs on the basis of testimony, our testimonial basis for our beliefs, being a posteriori, cannot be the only basis on which we hold them when we engage in reflective equilibrium reasoning. Our justification must rather be suitably a priori, tied to the weight that these beliefs pull or are expected to pull in our web of moral beliefs in reflective equilibrium. The requirement that we bring our judgements about specific cases and the general principles that we think likely justify them into a wide reflective equilibrium with the rest of our beliefs about the world should similarly be thought of as requiring us to square these with our judgements about other a priori truths, including the contingent a priori truth (if it is a truth) that we are capable of moral knowledge.

Suppose we have gone through this process, and that there seems to be no tension or conflict between our judgements about specific cases and general principles, and between these and other a priori truths. Does that entail that we have moral knowledge? No it doesn't. For one thing, there is always the possibility of our having made a mistake. What seems to us to make perfect sense

when we engage in reflective equilibrium reasoning might make no sense at all—what seems to us to be true a priori might be completely incoherent. For another, there is always the possibility of our having fallen victim to the garbage-in, garbage-out problem. If we start from premises that are so off-base that we couldn't reason ourselves to moral knowledge by starting from them, then we are doomed from the start—think of a die-hard skeptic who starts from skeptical premises. But if no such mistakes have been made, and if the skeptical possibility that the premises from which we start aren't ones from which we could reason ourselves to moral knowledge is ruled out, then, on the supposition that moral rationalism is true, moral knowledge is indeed the upshot.

It is worth comparing this account of the way in which we could come by knowledge of moral facts, on the supposition that moral rationalism is true, with Thomas Scanlon's suggestion that all existence claims associated with pure claims within a discourse, and in particular all pure claims within moral discourse, are domain-specific, where pure claims within a discourse are those that presuppose the truth of no claims in any other discourse, and where domains are individuated by the distinctive concepts and modes of argument that are employed within the discourse in which such claims are made to establish their truth (Scanlon 2014). There are both similarities to and differences from what has just been said about our knowledge of moral facts, assuming moral rationalism to be true, and what Scanlon says about our knowledge of the existence claims associated with pure claims within moral discourse.

According to Scanlon, if we reach the conclusion that pure claims within a discourse are justified by using the characteristic mode of argument within that discourse in support of those claims, then, in taking those claims to be justified, we thereby commit ourselves to supposing that the existence claims associated with those pure claims are likewise true. To be committed to the truth of pure claims within a discourse just is to be committed to the existence of the properties and relations required to make those claims true. In the case of pure claims within a discourse, it therefore follows that the only challenges to the associated existence claims that can arise are challenges internal to the mode of discourse itself—this is what it means for such claims to be domain-specific. Scanlon is especially interested in the upshot of this for pure *moral* claims, given that as he sees things pure moral claims commit us to the truth of claims about the existence of an irreducible reason-relation—more on this in the next section.

Let's begin with the similarities between what we have said and what Scanlon says. To put our point in Scanlon's terms, the domain in question when we figure out what the moral facts are is the domain of moral discourse; the claims in question are claims about the existence of non-moral kinds of acts that are, in virtue of their being the non-moral kinds they are, morally obligatory, morally permissible, morally forbidden, and supererogatory; the characteristic mode of argument employed within that mode of discourse to establish the truth of such claims is reflective equilibrium reasoning; and the suggestion has been that so long as we are in a position to properly employ that mode of argument and do in fact employ it properly, we thereby gain a priori moral knowledge. So far, so good. The question, however, is whether we should suppose that these judgements about non-moral kinds of acts that are, in virtue of their being the non-moral kinds they are, morally obligatory, morally permissible, morally forbidden, and supererogatory, are *pure* claims within moral discourse.

For these to be pure claims within moral discourse it would need to be the case that the requirement that we bring our judgements about specific cases and general principles into a wide reflective equilibrium with the rest of our beliefs about the world has no real bite. But is that so?

Are there any a priori truths, truths that aren't themselves moral truths, with which we have to have been able to square our judgements about specific hypothetical cases and general principles in order to be justified in believing them? This question is particularly pressing for moral rationalists, as they hold that moral facts entail facts about reasons for action. That entailment is itself supposed to be a priori and necessary, so the question is whether there are two distinct domains of discourse, the moral domain and the domain of reasons, or just one. Or, to put the question slightly differently, in answering moral questions and questions about reasons for action do we draw on the same concepts and modes of argument, or are the concepts and modes of argument we draw on slightly different from each other?

Suppose the answer is that they are slightly different from each other. In that case moral claims are not pure claims, but rather presuppose the truth of claims in the different domain of discourse about reasons for action. In order to be justified in our moral beliefs, we would have to square our moral beliefs with our more or less independent beliefs about what reasons for action there are—I say 'more or less' because in wide reflective equilibrium our beliefs all depend on each other. Having convinced ourselves that it is (say) morally permissible to ϕ, we would have to confirm that there are indeed impartial and unconditional reasons for action, and that none of these are decisive reasons not to ϕ. In this way, our moral beliefs would be vulnerable to external challenge, as any independent skepticism we might have about the existence of impartial and unconditional reasons for action would immediately lead to skepticism about morality. The question is whether a moral rationalist can resist thinking that our moral beliefs are vulnerable to an external challenge of this kind.

It might be thought that they can if they insist that there are not two distinct domains. Instead, they might say, there is just one domain, the domain of reasons for action, and the moral domain is simply a sub-domain within the domain of reasons for action. The most straightforward way in which to support this claim would be by insisting that the connection between moral facts and reasons for action is analytic. To judge that the various moral facts obtain, they might say, just is to judge that there are the corresponding reasons for action, where the difference between these reasons for action and others is the nature of the reason-giving features. Moral reasons for action are those in which the reason-giving features are impartial and unconditional, not so non-moral reasons for action.

This sounds like a promising strategy, but as with all promising strategies, the devil is in the details. The crucial question is how we are to characterize the domain of reasons for action so as to make this purported difference between moral and non-moral reasons for action come out true. More precisely, the question is whether the domain of reasons for action, properly understood, introduces its own distinctive concepts and modes of argument.

3. The Reasons-First View

One answer to these questions, the 'Reasons-First View', is inspired by Thomas M. Scanlon's work on reasons (1998, 2014). The basic idea behind this view is that there is a primitive reason-relation in terms of which all other normative features can be explained. The Reasons-First View is thus committed to both reasons primitivism (the claim there is a primitive reason-relation) and reasons fundamentalism (the claim all other normative features apart from the reason-relation can be explained in terms of the reason-relation).

Let's begin with reasons primitivism. According to Scanlon, the domain of reasons "in the standard normative sense" is a domain of discourse about a primitive four-place relation that

relates considerations, attitudes, persons, and circumstances. This relation is irreducibly normative, and it allows us to define a distinctive class of attitudes, the *judgement-sensitive attitudes*, where these are those attitudes . . .

> . . . that an ideally rational person would come to have whenever that person judged there to be sufficient reasons for them, and that would, in an ideally rational person, 'extinguish' when that person judged them not to be supported by reasons of the appropriate kind (1998: 20).

The reasons that Scanlon speaks of in this passage are the *considerations* in the four-place reason relation, and what these considerations provide are sufficient reasons in the standard normative sense for the relevant *attitudes* of the relevant *person* in the relevant *circumstances*. Considerations thus get to be reasons in virtue of their place in the four-place reason relation, and an ideally rational person is simply someone who is maximally sensitive to what they take such considerations to be in the formation of their judgement-sensitive attitudes.

The identification of the members of the class of judgment-sensitive attitudes is important, according to Scanlon, because these attitudes "constitute the class of things for which reasons in the standard normative sense can be asked or offered" (1998: 21). The paradigmatic example of such considerations are those that support the truth of our beliefs, so beliefs are judgement-sensitive attitudes *par excellence*, according to Scanlon. This is because it is in the nature of beliefs to be sensitive to the considerations that believers take to provide reasons for them. But there are other judgement-sensitive attitudes as well. These include attitudes like intention, desire, fear, and admiration. Scanlon insists that it is in the nature of all of these attitudes to come and go in an ideally rational person depending on what that person takes to be reasons for forming or ridding themselves of them.

An important feature of Scanlon's view, as already mentioned, is that the reason-relation itself is primitive. It is primitive because, as he sees things, we cannot explain that relation in other terms.

> Any attempt to explain what it is to be a reason for something seems to me to lead back to the same idea: a consideration that counts in favor of it. "Counts in favor how?" one might ask. "By providing a reason for it" seems to be the only answer (1998: 17).

As Scanlon immediately emphasizes, however, there are other normative features apart from the reason-relation, and the attraction of the reason-relation is that it can be used to explain these other normative features. This is where his reasons fundamentalism comes in. According to reasons fundamentalism, all normative features apart from the reason-relation inherit their normative status from their connection with the reason-relation.

For example, since actions are not attitudes, Scanlon holds that reasons for action are reasons in a different but related sense to the sense in which there are reasons for the judgement-sensitive attitudes. Considerations that provide reasons for actions provide them in virtue of the connection between actions and some judgment-sensitive attitude or other. Scanlon's preferred candidate is intention. As he puts it,

> . . . 'reason for action' is not to be contrasted with 'reason for intending'. The connection to action, which is essential to intentions, determines the kinds of reasons that are appropriate for them, but it is the connection with judgment-sensitive attitudes that makes events actions, and hence the kind of things for which reasons can sensibly be asked for

and offered at all (1998: 21).

Reasons for actions thus count in favor of actions by counting in favor of the intentions that produce those actions.

The upshot is that, even if we grant that there is a domain of reasons of the kind Scanlon describes, reasons for action are not basic elements in this domain. The basic elements in the domain are rather reasons for the judgement-sensitive attitudes. Reasons for action get explained by their link to reasons for the specific judgement-sensitive attitude of intending. Reasons for action are reasons for the intentions that produce them. And what goes for reasons for action goes for a whole range of other normative features as well. Intrinsic desirability turns out to be a matter of there being reasons to intrinsically desire—we will return to this example below; danger turns out to be a matter of there being reasons to fear; admirability turns out to be a matter of there being reasons to admire; and so on. None of these normative features are basic elements in a domain, but are instead elements within the domain of reasons because of their connection to the specific judgement-sensitive attitudes.

(Scanlon tells us that we can explain reasons for action in terms of reasons for intending because, on the one hand, it is in the nature of intentions to produce actions, and, on the other, it is in the nature of actions to be produced by intentions. There is a problem here, however, as there is no such link between actions and intentions—or, more precisely, there is only such a link on a very weak understanding of intentions according to which any desire or pro-attitude that can motivate an action counts as an intention. Many actions are, after all, produced by whims and fancies that don't have the various stability features associated with intentions. The link we need rather is the so-called "standard story" of action according to which what makes actions *actions* is the fact that they are produced by an agent's desires and beliefs (see Davidson 1963; Smith 1998, 2012). Reasons for action would then be explained by reasons for the desires and beliefs that produce those actions. A discussion of this point would, however, take us too far afield, as it would require us to engage with Scanlon's non-mainstream views about the nature of desires and their role in the production of action, so I will ignore the issue in what follows (but see Smith 2011).)

Consider now the connection between the Reasons-First View and moral rationalism. As should be clear, the Reasons-First View provides us with a more precise way in which to formulate moral rationalism, a way that makes the moral domain a sub-domain of the domain of reasons. Having explained what reasons for action are in terms of reasons for intentions, we then go on to explain what it is for actions to be morally obligatory, morally permissible, and morally forbidden in terms of there being certain sorts of reasons for action. In this way moral features too can be seen to inherit their normativity from the normativity of reasons for judgement-sensitive attitudes. Moreover, in virtue of these explanatory connections, moral epistemology turns out to be a part of the epistemology of reasons. There is no distinction between the modes of argument we employ in figuring out how the primitive four-place reason relation relates considerations, persons, judgement-sensitive attitudes, and circumstances, on the one hand, and what reasons for action there are, and which acts are obligatory, permissible, and forbidden, on the other. To have pure moral knowledge is already to have knowledge of reasons for action and reasons for intentions. The moral rationalist's claim that moral facts entail facts about reasons for action thus provides us with no grounds at all for supposing that moral knowledge is vulnerable to challenge from outside the moral domain because the moral domain is already contained within the domain of reasons.

What should we think of the Reasons-First View? Let's begin with the reasons primitivism. Scanlon tells us that if we ask what it means to say that a consideration is a reason, though we could say that it is a consideration that counts in favor, if we were to ask how it counts in favor, all we could say is that it counts in favor by being a reason. This is why he thinks that the feature is irreducibly normative. But it is evidently false that that's the only thing we could say. Suppose someone tells me that I should believe that the Earth is flat, and I ask them for a reason. In one scenario, they reply that they will give me $1 million if I believe that the Earth is flat. In another scenario, they reply by drawing my attention to the fact that the horizon line looks flat. Now suppose I ask this person how the consideration that they have provided counts in favor. In neither scenario would we expect them to respond by saying that it counts in favor by being a reason.

In the first scenario in which I ask them how their giving me $1 million counts in favor of my believing that the Earth is flat, what we would expect them to say is that it counts in favor by making it desirable for me to believe that the Earth is flat, and we would then expect them to explain what the relevant desirability characteristic is. For example, they might think that benefits are desirable, and tell me all about the benefits of my being rich. But in the second scenario in which I ask them how the fact that the horizon line looks flat counts in favor of my believing that the Earth is flat, we would expect them to say something completely different. We would expect them to say something along the lines of things generally being the way they look, and hence that the flat look of the horizon line counts favor of believing that the Earth is flat by supporting the truth of that proposition.

What we have here are thus two very different ways of spelling out what it is for a consideration to count in favor of a belief, but neither averts back to the ambiguous claim that what's said to count in favor does so by being a reason to believe that the Earth is flat. What was requested and given is instead a disambiguation of that ambiguous claim. Moreover, only one of the ways of disambiguating what it is for a consideration to count in favor of believing is plausibly a way of spelling out what it is for a consideration to count in favor by way of being a reason in the "standard normative sense" for believing, where it is in the nature of beliefs to be acquired and given up by an ideally rational person depending on whether they judge there to be sufficient reason for believing in that sense. For though it is indeed in the nature of belief to be sensitive to considerations that are taken support their truth—what it is to be a belief is inter alia to be a state that comes and goes in response to considerations of truth-conduciveness—it is not in the nature of belief to be sensitive to considerations concerning their desirability. That beliefs are sensitive to considerations concerning their desirability, if they are, is at best a contingent and a posteriori fact about belief.

With this distinction in mind, consider now reasons for action. Suppose someone tells me that I should go home immediately, that I ask them for a reason, and that they tell me that my wife needs my help. If I ask them how my wife's needing my help counts in favor of my going home, what we would expect them to provide is an account of the feature possessed by my going home when my wife needs my help that makes my going home choiceworthy or desirable. For example, they might tell me that if I go home then I could provide my wife with the help that she needs. Reasons for action thus seem to be reasons in the sense of 'reason' that has nothing to do with reasons in the standard normative sense. They are more like reasons for believing that show believing to be desirable, and nothing at all like reasons for believing that show believing to be supported by considerations that support the truth of what is believed.

On the face of it, this is very bad news for the Reasons-First View. It is very bad news because it undermines both reasons primitivism and reasons fundamentalism. It undermines reasons primitivism because it suggests an alternative reductive explanation of the normativity of reasons. On this alternative explanation, the normativity of reasons is explained in the first instance—more on this in the next section—by the functional nature of belief, where this is in turn spelled out in terms of the notion of truth-conduciveness. It is in the nature of belief to be a state that has a certain functional role, where a specification of this functional role has two parts. One part concerns the relationship between belief and the world. Beliefs that function optimally constitute knowledge, so optimally functioning beliefs must match the world—that is, they must be true—and where this matching is itself a product of regulation by the world, that regulation must itself be non-accidental. The other part concerns the characteristic role that beliefs play in our psychological economy. In the case of optimally functioning beliefs, that role is a matter of their coming or going depending on whether what is believed is supported by considerations the subject takes to be conducive to the truth of what is believed.

The attraction of this reductive explanation of the normativity of reasons is that it turns out to be an instance of a more general kind of functional normativity (compare Smith forthcoming). Think of functional kinds like hearts. The function of the heart is to pump an adequate supply of blood around the body, from which it follows as a matter of definition that someone's heart ought to pump an adequate supply of blood around the body. This follows as a matter of definition because, quite in general, there is a sense of 'ought' in which things of a functional kind ought to function in the way that optimally functioning things of that kind do function. But if this is right then the claim that someone ought to believe what's supported by what they take to be reasons turns out to be an instance of this more general truth. Belief is a functional kind too, and beliefs that function optimally come and go depending on whether the believer takes there to be considerations that support the truth of what's believed, where reasons for belief just are those truth-supporting considerations. The details would no doubt be messy and difficult to spell out in detail, but the basic idea should be clear enough.

It might be replied on Scanlon's behalf that, even without getting into the details, we can already tell that this reductive explanation of the normativity of reasons is inadequate, and that Scanlon's reasons primitivism is preferable. The only psychological states for which truth-conduciveness so much as makes sense are those that either are, or have as parts, states like belief that can be true or false. It might therefore be thought to be an implication of the proposed reductive account of the reason-relation that reasons can only be given for this very narrow range of psychological states. But, the objection goes, there are plainly psychological states for which there are reasons in the standard normative sense that aren't like this. Intention is an obvious example, intrinsic desire is another, fear is another, admiration is another, and so on. Focus on the case of intrinsic desire.

An intrinsic desire to avoid scratching one's finger even at the cost of the destruction of the whole world is a psychological state that cannot itself be true or false, and nor does it have as a part a psychological state that can be true or false. But it is still a psychological state there is a decisive reason in the standard normative sense not to have. The best explanation of this, the objection continues, is that offered by reasons primitivism. There is a consideration—the pain and suffering that would ensue from the destruction of the whole world by comparison with the minor inconvenience associated with scratching one's finger when one doesn't want to—that

counts in favor of not having such an intrinsic desire. The reductive proposal is thus objectionable on purely extensional grounds, or so the objection alleges.

The objection misunderstands the reductive proposal. What makes the considerations that support the truth of beliefs reasons for beliefs is not the fact that beliefs can be true or false, but rather a more general condition that beliefs satisfy, a condition that is also satisfied by intentions, intrinsic desires, fear, admiration, and all the other judgement-sensitive attitudes. This condition is spelled out by Judith Jarvis Thomson in her *Normativity* (2008). Belief, intention, intrinsic desire, fear, admiration, and all the other judgement-sensitive attitudes are mental states that have correctness conditions, where a correctness condition can be thought of informally as a condition that makes being in those mental states especially apt—Thomson says "deserved"—given the nature of those states. (More will be said about how these correctness conditions are to be explained in the next section.) What reasons in the standard normative sense for being in a mental state are, if Thomson is right, are *considerations that support the truth of the propositions that express those state's correctness conditions*.

More slowly, just as the truth of the proposition believed is the correctness condition of belief—that is, the state of the world that makes it especially apt to believe that that is the state of the world, given the nature of belief—so the intrinsic desirability of what's intrinsically desired is the correctness condition of intrinsic desire (that is, the state of the world that makes it especially apt to desire the things that are intrinsically desirable); the dangerous nature of the objects of fear is the correctness condition of fear (that is, the state of the world that makes it especially apt to fear the dangerous things); and so on. Quite in general, the truth of these propositions about the correctness conditions of these states is what makes not just belief, but also intrinsic desire, fear, and so on especially apt.

According to Thomson, with this characterization of what it is for various mental states to have correctness conditions in place, we can then give the following quite general characterization of reasons in the standard normative sense for being in mental states with correctness conditions (2009: 131):

> A reason for being in a mental state with a correctness condition is a consideration that supports the truth of the proposition that is that mental state's correctness condition.

Belief thus turns out to be a state for which there are reasons in the standard normative sense not because beliefs can be true or false, but rather because the correctness condition of a belief is the truth of the proposition believed, and because reasons in the standard normative sense for believing are therefore considerations that support the truth of the proposition believed. Intrinsic desire is a state for which there are reasons in the standard normative sense for the same reason. The correctness condition of an intrinsic desire is the truth of the proposition that the object of the intrinsic desire is intrinsically desirable, and reasons in the standard normative sense for intrinsically desiring are therefore considerations that support the truth of the proposition that the object of the intrinsic desire is intrinsically desirabile. The same goes for each of the other judgement-sensitive attitudes.

Notwithstanding the fact that the reason-relation isn't primitive, we can therefore still characterize the judgement-sensitive attitudes in terms of reasons, and the judgement-sensitive attitudes include all of the psychological states that Scanlon mentions: belief, intrinsic desire, intention, fear, admiration, and so on. This is in turn important, as it shows that those who disagree with the Reasons-First View needn't disagree with them about the extension of reasons

for judgement-sensitive attitudes. There might well be all the reasons that those who advocate the Reasons-First View say there are. What those who reject the Reasons-First View take issue with is rather reasons primitivism or reasons fundamentalism, and what we have seen so far is that they have good reasons to reject reasons primitivism. The reply on Scanlon's behalf to the reductive proposal thus misses its mark.

It will be useful to have a name for Thomson's view about the way in which reasons connect up with the correctness conditions of the judgement-sensitive attitudes, so let's call it the Reasons-Correctness Nexus. The Reasons-Correctness Nexus is the claim all reasons "in the standard normative sense" are reasons for attitudes and are considerations that conduce to the truth of the correctness conditions of those attitudes. However, as is perhaps already clear, the Reasons-Correctness Nexus suggests that there is a further problem for the Reasons-First View as well. Given the Reasons-Correctness Nexus, and given the rejection of reasons primitivism, we have good reasons to reject reasons fundamentalism.

According to reasons fundamentalism, you will recall, we are supposed to be able to explain all normative features in terms of some judgement-sensitive attitude's connection with reasons. For example, what it is for p to be intrinsically desirable is supposed to be explained in terms of there being reasons to intrinsically desire that p. But p's being intrinsically desirable cannot be explained in terms of there being reasons to intrinsically desire that p if we have to explain what reasons to intrinsically desire that p are in terms of considerations that support the truth of the proposition that p is intrinsically desirable. To attempt to explain what reasons to intrinsically desire that p are in terms of considerations that support the truth of the claim that there are reasons to intrinsically desire that p would be to presuppose the very thing that we are trying to explain. Given the Reasons-Correctness Nexus, it therefore follows that being intrinsically desirable must rather be a feature whose nature can be explained independently of reasons. This is simply the denial of reasons fundamentalism.

To sum up, our aim in this section has been to explore the Reasons-First View, which is the view that all normative facts, except for facts about reasons for judgement-sensitive attitudes, are to be explained in terms of facts about reasons for judgement-sensitive attitudes which cannot themselves be explained at all. The Reasons-First view thus consists of two claims: reasons primitivism and reasons fundamentalism. We saw initially that we have good reasons to reject reasons primitivism in favor of an explanation of what reasons are in terms of truth-conduciveness, and we saw subsequently that we also have good reasons to reject reasons fundamentalism. Moral rationalists therefore have good reasons to reject the Reasons-First View.

4. The Desirability-First View

The question with which we began is whether a moral rationalist can resist the idea that there are two distinct domains, the domain of moral facts and the domain of facts about reasons for action. If they cannot resist this idea, then even if our moral beliefs about specific cases and the general principles that justify them were in reflective equilibrium with each other, these beliefs would still be vulnerable to external challenge on the basis of skepticism about the existence of corresponding reasons for action.

The answer to this question given by advocates of the Reasons-First View was that a moral rationalist can resist this idea. In their view, the moral domain is a sub-domain within the domain of reasons. If they had been right, then having our moral beliefs about specific cases and the general principles that justify them in reflective equilibrium with each other would already have

been to have our beliefs about corresponding reasons for action in reflective equilibrium with each other. But given that facts about reasons for action entail facts about desirability that themselves do not reduce to facts about reasons, it follows that this answer isn't available. But is a structurally similar answer available?

Suppose moral rationalists adopt the Desirability-First View. According to this view, inspired by G. E. Moore's *Principia Ethica* (1903), there is a primitive normative property of being intrinsically desirable and all other normative features are defined in terms of it. An agent has a reason to act in a certain way just in case his so acting would realize an intrinsically desirable outcome, and then we define being morally obligatory, being morally permissible, and being morally forbidden as before. There is thus a domain of facts about what is intrinsically desirable, the domain of facts about reasons for action is a sub-domain within this domain of facts, and the moral domain is a sub-domain within the domain of facts about reasons for action.

As with the Reasons-First View, the advocate of the Desirability-First View could suppose that the difference between moral and non-moral intrinsic desirability lies in the nature of the features in virtue of which things are intrinsically desirable. In the case of outcomes that are intrinsically desirable morally, they could say, the features in virtue of which they are intrinsically desirable are suitably impartial, and they make for intrinsic desirability unconditionally, not conditionally on the presence of the desires of those who brought about those outcomes. In the case of things that are intrinsically desirable non-morally, these conditions needn't be met. Moral epistemology thus turns out to be the epistemology of reasons for action, which turns out to be the epistemology of intrinsic desirability, and so once again pure moral knowledge becomes invulnerable to an external challenge.

Finally, given the Reasons-Correctness Nexus, the advocate of the Desirability-First View could insist that the very same reasons that provide us with a priori knowledge of facts about what is intrinsically desirable morally, knowledge we could gain a priori, also provide us with reasons to form intrinsic desires with contents that match those intrinsic desirability judgements. For example, if there are reasons that provide us with knowledge that happiness is intrinsically desirable, then those same reasons would support our intrinsically desiring happiness. Whenever we had reasons to believe that certain actions have intrinsically morally desirable outcomes, we would have corresponding reasons to have the intrinsic desires that would lead us to bring those outcomes about. The Reasons-Correctness Nexus could thus explain why the Desirability-First View is a version of moral rationalism.

How plausible is the Desirability-First View? The problem facing advocates of the Desirability-First View is to explain something that we have so far taken for granted. Why do correct desires have the desirability-making features of outcomes as their contents? Correct beliefs have true propositions as their contents, and the explanation of this draws on what we know about the functional nature of belief alluded to earlier. It is in the nature of belief not just to be sensitive to considerations that are taken to support their truth, but also to be a state that can reliably combine with desire so as to lead agents to act in ways that satisfy their desires. The only beliefs that are capable of doing this are those that constitute knowledge, and hence those whose contents are true. For though agents may sometimes satisfy their desires when they act on their false beliefs, or their accidentally true beliefs, their doing so is purely a matter of luck. This is why I said earlier that beliefs that function optimally constitute knowledge.

We can now state more precisely the problem faced by advocates of the Desirability-First View. The problem is that we need to be able to tell a similar story about why correct intrinsic desires have intrinsic desirability-making features of outcomes as their contents. To be relevantly similar, the story would have to draw on the nature of intrinsic desire and intrinsic desirability. It would have to be a story according to which it is in the nature of intrinsic desire to play some functional role or other, and having the intrinsic desirability-making-features of outcomes as their contents would have to be what enables them to play this role optimally. Moreover, if the Desirability-First View is correct, some of the intrinsic desires in question would have to have impartial and unconditional intrinsic desirability-making-features of outcomes as their contents. The question is what that functional role could be. Advocates of the Desirability-First View would seem to have only two options.

One role of intrinsic desires is a downstream role. Intrinsic desires with the intrinsic desirability-making features of outcomes as their contents combine with true beliefs so as to lead agents to act in ways that produce those outcomes. The advocate of the Desirability-First View might think that this role of intrinsic desires with intrinsic desirability-making features of outcomes as their contents dovetails with an independently plausible conception of an agent as someone whose nature is to bring about what's intrinsically desirable. But the trouble with this explanation is that it has little to do with the functional role of desire. To be sure, intrinsic desires do have the role of combining with true beliefs about how their contents are to be realized so as to realize their contents, but this is true independently of what the contents of those intrinsic desires are. Intrinsic desires with intrinsic desirability-making features and those with intrinsic undesirability-making features as their contents combine equally with true beliefs so as to realize their contents.

Nor does it help to talk of the 'independently plausible conception of an agent as someone whose nature is to bring about what's intrinsically desirable'. For even if we grant that that is indeed an independently plausible conception of an agent, what we are after is an explanation of why agents have this nature. What is it about intrinsic desirability and desire that makes this so? The case of belief is once again illustrative. It is, after all, independently plausible that an agent is someone whose nature isn't just to have beliefs about the world, or justified beliefs about the world, but knowledge of the world. But even though this is independently plausible, as we have seen, we can still explain why it is so by looking more closely at the functional role of belief. What we're after is a similarly compelling explanation of why it is in the nature of an agent to bring about what's intrinsically desirable by looking more closely at the functional role of intrinsic desire. What's needed is some connection between the functional role of intrinsic desire and intrinsic desirability.

Another role that advocates of the Desirability-First View think intrinsic desires play is an upstream role. Rational agents who lack intrinsic desires with intrinsic desirability-making features as their contents can come to acquire those intrinsic desires by engaging in reasoning about which outcomes have the primitive property of being intrinsically desirable. Finding reasons to believe that outcomes with (say) lots of happiness in them have the primitive property of being intrinsically desirable, they thereby find reasons to intrinsically desire outcomes with lots of happiness in them, and so, if they are sensitive to what they take these reasons to be, come to acquire such intrinsic desires and bring such outcomes about. They might think that this explains the independent plausibility of the idea that it is in the nature of an agent to bring about intrinsically desirable outcomes. But it doesn't explain that. For while it is true that intrinsic

desires could indeed be so acquired if correct intrinsic desires are those whose contents have the primitive property of intrinsic desirability, this 'explanation' presupposes that correct intrinsic desires have such contents, it doesn't explain why they do.

This shows us something important about the sort of explanation we need. We need an explanation of why correct intrinsic desires have intrinsic desirability-making features as their contents, and this explanation must appeal to something about the functional role of intrinsic desires, a role that goes beyond intrinsic desire's being such as to combine with true beliefs about how their contents are to be realized so as to realize that content. If moral rationalism can be given an adequate formulation at all, then something about this role must explain why intrinsic desires can be acquired through reasoning. Though we have not yet shown that advocates of the Desirability-First View can provide no such explanation, it has to be said that the prospects of their doing so look dim.

The problem for advocates of the Desirability-First View lies in their conception of intrinsic desirability as a primitive property. Think again about the case of belief. The reason we could explain why beliefs have true propositions as their content is because we could appeal to our platitudinous understanding of what it is for this to be so, namely, for the world to be the way it is believed to be. It is the fact that knowledge implies true belief that explains why agents who act so as to satisfy their intrinsic desires in the light of their knowledge end up satisfying their intrinsic desires. The problem for advocates of the Desirability-First View is that their primitivism prevents them from providing us with a similarly platitudinous understanding of what it is for an outcome to be intrinsically desirable. Moral rationalists therefore have good reason to abandon the Desirability-First View's primitivism about intrinsic desirability.

5. The Function-First View

It should be clear where we are headed. The Function-First View retains those aspects of the Desirability-First View that makes it superior to the Reasons-First View. It explains what moral facts are in terms of reasons for action, it explains what reasons for action are in terms of the intrinsic desirability of the outcomes of actions, and it explains what reasons for intrinsic desires are in terms of the Reasons-Correctness Nexus. However it rejects the Desirability-First View's commitment to primitivism about intrinsic desirability, holding instead that we can define what it is for the outcomes of an agent's actions to be intrinsically desirable in terms of what that agent's ideal counterpart intrinsically desires, where an agent's ideal counterpart is simply that agent in the nearest possible world in which she has beliefs and desires that function optimally (compare Smith 1994).

The full weight of this definition of intrinsic desirability is carried exactly where we saw it should be carried in our discussion of the Desirability-First View, namely, by the account we give of what it is for an agent's beliefs and desires to function optimally. We know that their functioning optimally means that an agent has knowledge of the world in which he lives, and we also know that if the world in which he lives is a world in which he has the option of bringing about the outcomes that he intrinsically desires, his knowledge of how to do that connects up with his intrinsic desires so as to bring about those outcomes. But what is at issue, when it comes to formulating moral rationalism, is whether this is an exhaustive characterization of what it is for his beliefs and desires to function optimally. If it is, then moral rationalism is doomed, as it would follow immediately that there are no moral reasons for action.

Moral reasons for action, remember, are characterized by the nature of the corresponding intrinsic desirability-making features, features which in turn are fixed, according to the Function-First View, by the contents of the intrinsic desires of an agent's ideal counterpart. There are moral reasons for action only if some of these intrinsic desires have contents that are impartial and unconditional. In other words, some of the intrinsic desires that agents have to have when their beliefs and desires function optimally must be intrinsic desires that have contents that are impartial and these intrinsic desires must be required for optimal functioning as such. In other words, agents' ideal counterparts must *converge* on intrinsic desires with these impartial contents (compare Smith 1994: 164–177). They must converge because only so would the corresponding intrinsic desirability-making features and reasons for action be unconditional. The problem with supposing that the characterization of what it is for beliefs and desires to function optimally given in the previous paragraph is exhaustive should now be clear. The problem is that, if it were exhaustive, then there would be no intrinsic desires with impartial contents that are required for optimal functioning as such. The intrinsic desires that agent's optimally functioning counterparts have would be a function of whatever intrinsic desires they happen to have.

We now have a clearer sense of what's required for moral rationalism to be true. There must be some hitherto unnoticed role for intrinsic desires to play in the optimal functioning of an agent's beliefs and desires. The following is an attempt to formulate an argument for just this conclusion inspired by some of the things that Kant says in the *Groundwork* (1785). The argument takes us from a specification of the functional role of intrinsic desires to the conclusion that all agents' ideal counterparts have certain intrinsic desires with an impartial content (note that I am not endorsing this argument—I give it for illustrative purposes only).

The argument begins with the observation that there is a sub-class of agents with the capacity to know which outcomes their ideal counterparts intrinsically desire, and to bring these outcomes about in the light of this knowledge. It then seeks to show that certain intrinsic desires must be possessed by those agent's ideal counterparts if agents are to have this capacity. Here is the argument.

1. Agents with the capacity to know which outcomes their ideal counterparts intrinsically desire to be brought about and to bring those outcomes about in the light of this knowledge have the capacity for self-governance.
2. For each agent with the capacity for self-governance, there is a possible world in which that agent exercises that capacity and so is self-governing.
3. For any arbitrary group of agents with the capacity for self-governance, there is a possible world in which those agents exercise that capacity and so are self-governing.
4. What agents with the capacity for self-governance do in those possible worlds in which they all succeed in being self-governing as part of an arbitrary group of such agents is, first, not interfere with what any other self-governing agent is doing, and second, not sit idly by while other agents fail fully to develop the capacity for self-governance or lose that capacity, but instead help them acquire or maintain it.
5. The only way that agents with the capacity for self-governance could act in these ways in those worlds is by having ideal counterparts with an intrinsic desire to help agents with the capacity for self-governance fully acquire and maintain these capacities, and

an intrinsic desire not to interfere with the exercise of these capacities by those who have them—for short, intrinsic desires to help and not interfere.

6. The ideal counterparts of all agents with the capacity for self-governance have intrinsic desires to help and not interfere.

What should we make of this argument?

The first thing to say is that this argument purports to do exactly what we have seen is needed. It purports to identify a hitherto unnoticed role for intrinsic desires to play in the psychology of an agent whose beliefs and desires are functioning optimally, namely, that of explaining how it is possible for agents to be self-governing. The claim is that for agents to be self-governing is for them to have the capacity to know what they would intrinsically desire to bring about if their beliefs and desires were functioning optimally, and to bring those outcomes about in the light of that knowledge, and that the only way they could do this, given (1)–(6), is if the beliefs and desires of their properly functioning counterparts include intrinsic desires to help and not interfere with other such agents. These intrinsic desires are impartial and unconditional, so they could explain why there are intrinsic desirability-making features and reasons for action that are impartial and unconditional, and hence why there are moral reasons, and these moral reasons could explain why certain actions are morally obligatory, morally permissible, morally forbidden, and supererogatory. Moreover, the explanation parallels quite closely the earlier explanation of why correct beliefs have true propositions as their contents. In both cases, correct psychological states—true beliefs, in the one case, and intrinsic desires with contents that are impartial and unconditional, in the other—are crucial to the very possibility of the psychological states in question functioning optimally.

The second thing to say is that in doing exactly what was needed, the argument purports to establish what's morally obligatory, morally permissible, morally forbidden, and supererogatory on grounds that are totally independent of what might be established by a reflective equilibrium argument that starts from our confident judgements about what's morally obligatory, morally permissible, morally forbidden, and supererogatory, and then seeks to find general principles that explain why those judgements are true. The mode of argument just given is entirely different. It begins from a claim about what's required for agents who are self-governing to have beliefs and desires that function optimally. In Scanlon's terms, the difference between these modes of argument shows that there are two distinct domains, the domain of facts about what's morally obligatory, permissible, forbidden, and supererogatory, and the domain of facts about psychological states functioning optimally. Unsurprisingly, it is therefore possible for the deliverances of these modes of argument to conflict, and the question, if they conflict, is which we should give more weight to in figuring what's morally obligatory, morally permissible, morally forbidden, and supererogatory. Since only one of these arguments would establish a connection between what's morally obligatory, morally permissible, morally forbidden, and supererogatory and the existence of reasons for action, namely, the one that gives more weight to the conclusion of the argument from the optimal functioning of beliefs and desires, it seem fairly clear which one we should give more weight to.

The third and final thing to say about the argument from (1)–(6) is that, unfortunately, it doesn't succeed as it stands. There is a crucial ambiguity in (3). Read in one way, it is a reiteration of (2). On this way of reading (3) it says:

(a) (x)(If x has the capacity for self-governance, then ((∃w)(x is self-governing in w)))

That is, for each agent with the capacity for self-governance, there is a possible world in which that agent is self-governing. Read in the other way, the way required to support the move to (4), it says:

(b) (x)(y)(If x and y have the capacity for self-governance, then ((∃w)(x and y are both self-governing in w)))

That is, for arbitrary groups of agents with the capacity for self-governance, there is a possible world in which they are all self-governing. Moreover there is no way to move from (a) to (b) without making further assumptions about what's required to be self-governing. Imagine, just to keep things simple, an agent who has just one intrinsic desire, and that that is an intrinsic desire to interfere with some other agent's exercise of their capacity for self-governance. Such an agent is not a counterexample to (a), but is a counterexample to (b), as in the possible worlds in which she succeeds in being self-governing, some other agent does not succeed in being self-governing. This is not to say that there isn't some further argument that licenses the move from (a) to (b). The point is simply that until we have such an argument, we don't have an argument for the existence of moral reasons. But though the argument from (1)–(6) doesn't succeed as it stands, the fact that we have been able to spell it out to the point where we can identify the further argument that needs to be given is suggestive. Perhaps (1)–(6) could be fixed up, or perhaps there is some other argument.

To sum up, in any version of the Function-First View, an important role will be played by an argument along the lines of (1)–(6). Such an argument will begin from a claim about a hitherto unnoticed role played by intrinsic desires in the psychology of an agent whose beliefs and desires are functioning optimally. Different versions of the Function-First View will then be more or less plausible depending on the plausibility of the claim they make about the function of intrinsic desires and the argument they go on to give starting from this claim. In defending some version of the Function-First View, moral rationalists should therefore put all of their effort into coming up with a plausible version of an argument along the lines of (1)–(6) (Korsgaard 2009 makes such an attempt). The aim of the argument, to repeat, is to derive the substance of our moral reasons from a role that it is in the nature of intrinsic desires to play. Armed with such an argument, moral rationalists would be in a position to put morality on a rock solid foundation. Without such an argument, they must either suspend judgement on what moral reasons for action there are, assuming they remain confident that some such argument is forthcoming, or, if they lose confidence that some such argument is forthcoming, they must deny that there are any moral reasons (note that I here resolve an ambivalence at the end of Smith 2010).

6. Moral epistemology for moral rationalists—again

At the outset I noted that many contemporary moral rationalists hold that our moral beliefs are immune to external challenge. I also noted that whether they are right about this depends on how moral rationalism is best formulated. We saw above that there are two ways of in which we could formulate moral rationalism, the Reasons-First View and the Desirability-First View, that support a conception of moral beliefs as immune to external challenge. However both of these turn out to be views that we have independent reasons to reject. The question is whether the best formulation of moral rationalism, the Function-First View, similarly supports the idea that our moral beliefs are immune to external challenge, and the answer is that it doesn't.

As we have seen, the Function-First View suggests a rather different picture. There are two paths to moral knowledge, and the seamless path to moral knowledge therefore turns on those paths

leading to the same destination. We go down one path when we start with our beliefs about what's morally obligatory, morally permitted, morally forbidden, and supererogatory and try to get these into a reflective equilibrium with each other. If we succeed, we thereby commit ourselves to the truth of the various moral claims that are the objects of our moral beliefs, but we don't have moral knowledge until we square our commitment to these moral claims with the existence of corresponding moral reasons for action. We go down the other path when we begin with what we know about reasons for action—that is, that they reduce to facts about what our ideal counterparts intrinsically desire us to do—and then attempt to figure out, just from this premise, what substantive reasons for action we have, and whether any of these reasons are impartial and unconditional, by coming up with an argument along the lines of (1)–(6).

We can put the same point in Scanlon's terms. Given that we individuate domains of facts by the kinds of argument that are characteristically given in establishing the truth of claims within that domain, moral rationalism, when properly formulated, suggests that there is no domain of pure moral facts. All moral claims have implications for what is going on in the rather different domain of facts about the functional natures of our psychological states. The moral domain is the domain of facts about what we are morally obliged, morally permitted, and morally forbidden to do, which can be thought of as the same domain as that of facts about reasons for action, which can be thought of the same domain as that of facts about the outcomes of agents' actions being desirable, which can be thought of as the same domain as that of facts about the outcomes that agents' ideal counterparts desire, where being ideal is spelled out in terms of optimal functioning. Ordinary moral reasoning commits us to the truth of claims of each of these kinds. But the last of these commits us to truths that we establish by using a rather different style of argument. This is the domain of facts where we get much more specific about the natures of psychological states, and what we are required to do in establishing truths within this domain is to come up with convincing arguments along the lines of (1)–(6).

Perhaps unsurprisingly, it seems to me this is why both Mackie and Joyce think that we should be error theorists about morality. In their view, though ordinary moral reasoning commits us to the truth of various claims about what we are morally obliged, morally permitted, and morally forbidden to do, when we look at what we would have to be able to demonstrate about the functional natures of our psychological states for these moral claims to be true, we discover that no such demonstrations are forthcoming, and this undermines our moral commitments. Though I disagree with them on this last point—I think that such demonstrations are forthcoming (see for example Smith 2013)—I agree with Mackie and Joyce that our moral commitments are hostage to the possibility of such demonstrations. One of the great virtues of the Function-First View is that it makes it vivid why this is so.

REFERENCES

Dworkin, R. 1996. Objectivity and Truth: You'd Better Believe it. *Philosophy and Public Affairs* 25: 87–139.

Joyce, R. 2002. *The Myth of Morality*. Cambridge: Cambridge University Press.

Kant, I. 1785. *Groundwork of the Metaphysics of Morals*.
Korsgaard, Christine 2009: *Self-Constitution: Agency, Identity, and Integrity* (Oxford: Oxford University Press)
Mackie, J. 1977. *Ethics: Inventing Right and Wrong*. Harmondsworth: Penguin.
McPherson, T. 2008. Metaethics and the Autonomy of Morality. *Philosophers' Imprint* 8: 1–16.
———. 2011. Against Quietist Normative Realism. *Philosophical Studies* 154: 223–40.
Moore, G. E. 1903. *Principia Ethica*. Cambridge: Cambridge University Press.
Portmore, D. 2011. *Commonsense Consequentialism: Wherein Morality Meets Rationality*. New York: Oxford University Press.
Rawls, J. 1951. Outline of a Decision Procedure for Ethics. *Philosophical Review* 60: 177–97.
———. 1971. *A Theory of Justice*. Cambridge, MA: Harvard University Press.
———. 1974. The Independence of Moral Theory. *Proceedings and Addresses of the American Philosophical Association* 47: 5–22.
Scanlon, T. M. 1998. *What We Owe to Each Other*. Cambridge, MA: Harvard University Press.
———. 2013. *Being Realistic about Reasons*. Oxford: Oxford University Press.
Smith, M. 1994. *The Moral Problem*. Oxford and Cambridge, MA: Wiley-Blackwell.
———. 1998. The Possibility of Philosophy of Action. In *Human Action, Deliberation and Causation*, ed. J. Bransen and S. Cuypers, 17–41. Dordrecht: Kluwer Academic Publishers.
———. 2010. Beyond the Error Theory. In *A World Without Values: Essays on John Mackie's Moral Error Theory*, R. Joyce and S. Kirchin, 119–139. New York: Springer.
———. 2011. Scanlon on Desire and the Explanation of Action. In *Reasons and Recognition: Essays on the Philosophy of T.M. Scanlon*, ed. R. K. S. Freeman and R. J. Wallace, 79–97. New York: Oxford University Press.
———. 2012. Four Objections to the Standard Story of Action (and Four Replies). *Philosophical Issues* 22: 387–401.
———. 2013. A Constitutivist Theory of Reasons: Its Promise and Parts. *LEAP: Law, Ethics, and Philosophy* 1: 9–30.
———. forthcoming. Constitutivism. In *Routledge Handbook of Metaethics*, ed. T. McPherson and D. Plunkett. London: Routledge.
Thomson, J. J. 2008. *Normativity*. Chicago: Open Court Publishing Company.
Wolf, S. 1982. Moral Saints. *Journal of Philosophy* 79: 419–39.