1

What would an adequate theory of rationality be like?

THE FUNDAMENTAL NORMATIVE NOTION

When we argue with other people about what to do, very often we appeal to principles. Certainly when philosophers offer moral theories, and argue that we should be moral, they appeal to principles. And even when we, or they, offer reasons in place of principles, it is reasonable to think of such arguments as shorthand for appeals to principles. For no one would advocate an action simply because there was some reason in its favor, if it were clear that there were compelling reasons against performing it. Thus when reasons are cited in arguments, there is some idea that all the relevant reasons, taken together, support the action. This implies that there is some principle in the background that produces overall verdicts based on all those reasons: perhaps it is the simple principle ‘perform the action supported by the most reasons’, or perhaps it is some more complicated principle. One cites particular reasons in order to suggest that those reasons are sufficient to determine the outcome of the application of such a principle. The very plausible idea that two actions to which the same reasons are relevant must have the same rational status also suggests that reason-based arguments are backed by a unique principle: a principle that takes those reasons as input and yields the status of the action as output.

When a principle is made explicit in an argument, it is often appropriate to ask ‘Why should I follow that principle?’ And when an answer is given, in terms of some other principle, it is often appropriate to ask exactly the same question. In some cases there will be no good answer to this question, and then the recommendations that flow from the principle may lose their authority. There is, however, a significant philosophical tradition according to which this sequence of principles and questions, and more basic principles and further questions, cannot go on forever. At some point, after the articulation of one of these principles, it will no longer make sense, or be appropriate, to ask ‘But why should I follow
that principle?’ That is, there is a philosophical tradition that asserts the existence of a fundamental normative principle applicable to action. Perhaps this tradition goes back as far as Aristotle, who asserted that there was one governing end according to which all human action was to be judged. Hume, when he argued that reason cannot by itself direct the will, was reacting against the majority opinion of his contemporaries, according to whom reason could do so. That is, the philosophers against whom Hume was arguing held that if it could be shown that reason required or prohibited an action, that was the end of the practical argument about that action: no further appeal could possibly be made that could legitimately alter such a judgment. Kant also is a prominent member of this tradition, advocating the existence of a categorical imperative that tells one how one must act, and against which no further consideration can have any legitimate force.

Contemporary philosophers also defend the existence of a fundamental normative principle, or set of principles. Indeed, this is the sense of ‘rational’ that is central to contemporary ethical theorizing. For example, Stephen Darwall writes that “It is part of the very idea of the [rationally normative system] that its norms are finally authoritative in settling questions of what to do.” Thomas Nagel writes that it should not be possible to ask why one should do what one has reason to do, and that for this reason there cannot be a justification for acting rationally. And Allan Gibbard’s notion of rationality “settles what to do . . . what to believe, and . . . how to feel.”¹ According to all of these philosophers it is a conceptual truth that there cannot be a sufficient reason to act irrationally and that there is a reason not to do so. Therefore, according to these philosophers, the question ‘Could I have a sufficient reason to do an irrational act?’ is as misguided (or trivial) as the question ‘Could there be an unmarried bachelor?’

When we are presented with any proposal regarding this fundamental normative principle, there are two tests we can apply to see whether it is adequate. The first is to see whether the question ‘Why should I always follow that principle?’ makes clear sense. If it does make sense, this casts the fundamental nature of the principle into doubt. For the

¹ Darwall (1983), pp. 215–16; Nagel (1970), pp. 1–9; Gibbard (1990), esp. p. 49. See also Korsgaard (1996a), p. 104 and Smith (1994), pp. 150ff. Smith claims that it is all and only reasons which spring from the norms of rationality that make actions desirable. Of course there are other conceptions of rationality. Robert Nozick (1993), pp. 46, 117 is concerned with the human faculty of rationality, and is content to ask about its purpose or function.
principle is supposed to be the most basic – the principle that stands behind all others. If the above question makes sense, then the putative fundamental principle certainly is not wearing its fundamental nature on its face. That is, it does not appear to be the end of the normative road. The second test is to see whether one could ever sensibly offer reasons for acting against the principle. If this is a real possibility then the principle cannot be the fundamental principle that tells us how we ought always to act. To illustrate these tests, it may be useful to use them to disqualify one possible fundamental normative principle: always act so as to maximize the satisfaction of your preferences. Does it make sense to ask ‘But why should I always act so as to maximize the satisfaction of my preferences?’ Yes, it does. For one could elaborate the question in this way: ‘Why should I always act so as to maximize the satisfaction of my preferences, if I know my preferences are the result of a brain defect that tends to produce self-destructive preferences?’ This failure to pass the first test is related to the way in which the proposed principle will also fail the second test. For one way of sensibly offering a reason to act against the principle is to say ‘But if you follow this principle you will cause yourself a lot of pain, without any benefit.’

The second of the above tests is quite clearly one which a fundamental normative notion must pass. If there can be an adequate reason to act against a principle, that principle cannot be telling us how we ought always to be acting. The first test, however, is more slippery, and it may be useful to show how a principle may pass it without at first seeming to do so. Consider then the following:

One should never perform an action that will harm oneself unless it will bring some compensating benefit to someone (perhaps oneself). All other actions are rationally permitted.

It seems obvious that one could sensibly ask ‘Why should I always follow this principle?’ One reason it seems obvious is that there seems to be an answer. For example, one might offer ‘Because then one will avoid

2 It is unclear if any contemporary philosophers hold such a simple version of this view. But the criticisms offered here also tell against more sophisticated versions of such principles. For a very clear presentation of these criticisms see Ripstein (2001).

3 This is not the place to descend into arguments about the various ways in which one might patch up the suggested principle. But it is worth mentioning that the strategy of ruling out desires that are the result of, say, a brain defect, is not a simple one. For we use the notion of rational action in determining what counts as a brain defect, rather than (say) a statistically rare configuration of neurons.
suffering harms.’ But in fact that is not an answer, since it is false that if one successfully follows this principle, one will necessarily avoid suffering harms. This is because the principle permits one to suffer harms in cases in which one will thereby produce compensating benefits for someone else. One might then suggest the following amended answer: ‘Because one will avoid suffering harms, except in cases in which one will thereby produce compensating benefits for someone.’ What is important to see is that with this amended answer one has ceased to offer a further reason to obey the principle. One has simply pointed out that by following the principle one follows the principle. Of course, this brief discussion has not shown that the above principle actually does pass the first test. It only shows one way in which a principle may misleadingly appear to fail it. Moreover, though the above principle may in fact pass this one particular test, it may be inadequate for other reasons.

This book is part of the tradition that seeks to discover and defend a fundamental normative principle applicable to action – of course by some means other than the production of a still more fundamental principle. That is, it seeks to provide an account of a principle that passes both of the tests mentioned above. It is devoted entirely to this principle, and not to its employment in arguing for further normative claims. In particular, no moral view is advocated, although it will be clear that the account has significant implications for the development of moral views.

In the phrase ‘fundamental normative principle,’ the word ‘fundamental’ should not be taken to mean ‘most important.’ For there are many other normative principles that, in different contexts, are likely to be more important and more salient than the principle that is the central topic of this book. Of course, we should never follow these more salient principles if they can be shown to violate the fundamental one: that is part of what it means for it to be fundamental. Another part of what it means is that if it is clear that an action does not violate the fundamental principle then there may be nothing we can say to dissuade even a rational agent from performing it – the agent may remain perfectly rational in resisting all our arguments. As will become clear later in the book, this means that the fundamental normative principle gives agents a very wide scope in making decisions about how to act. Because of the lack of guidance that the principle provides, some may be tempted to think that it cannot really be fundamental. But that is to confuse being fundamental with being most generally useful, or with being salient. It will turn out that, because we almost always act rationally without having to think
Conditions on an adequate theory

about it, the fundamental normative principle will very rarely be of much use in particular decisions. It will not tell us, for example, which career to choose, or whether to marry, or to have children, or whether to pursue wealth over enlightenment. As I will argue in various ways in what follows, it will not even tell us whether to take the high moral road, or the low. These questions we must answer for ourselves – they are choices, and it is futile to search for a basic principle that will authoritatively hand us the correct answer. In a limited number of cases I have found that when people are tempted to act against the fundamental normative principle, it is sometimes effective simply to point this out. This tends to bring the real source of the temptation into clearer focus, which helps in resisting it. But the primary usefulness of a clear view of the fundamental normative principle is not – at least directly – practical. Rather, it is theoretical: the principle will figure in an explanation of what it is for an action to be rational, in a sense that is closely connected with mental functioning. This notion, in turn, is often indispensable in restricting the scope of ‘everyone’ as it is used in philosophical theories (such as contractualism). The principle will also play a role in explaining why we should want to be rational, in that sense. And of course the fundamental principle will have many indirect practical implications, for very often such a principle plays an obvious and central role in the development of moral theory. And a moral theory, if it is clear, may have significant practical implications for people who care about morality.

RATIONALITY AND MENTAL FUNCTIONING

It seems fairly clear that whatever the fundamental normative notion might be, it will use the facts about one’s situation in yielding its judgments. This is why, when we are trying to decide how to act, we do not simply rest content with our present beliefs or evidence about the consequences of our actions, but seek out additional relevant information. Seeking this information is part of the process of figuring out what to do. Sometimes, through no fault of our own, we may fail to get the correct information, or may form justified, but false, beliefs. Because of this we may often fail to discover what we ought to do, and consequently we may fail to do what we ought to do. In failing to act according to the fundamental normative principle in such cases, we are not to be blamed. Nothing has gone wrong in the mental processes that produced our action. We would not want to call such actions ‘irrational,’ if we were taking irrational action to count
against the rationality of the agent in a way that was relevant to questions of moral responsibility, competence to give consent, freedom of will, mental health, and so on. Similarly, we may sometimes perform an action that is permitted according to the fundamental normative principle, given the facts; however, given our beliefs, it may be that our performance is obviously the result of some mental malfunction. In such cases we may want to call the action ‘irrational,’ if we are concerned with these same questions of moral responsibility, competence, and so on.

Since there may often be adequate (but unknown) reasons to perform actions that would be irrational in this ‘mental functioning’ sense, it should be clear that the ‘mental functioning’ sense of rationality is not the fundamental normative sense. It fails the second test. Nevertheless, it should be equally clear that the two senses of rationality are very closely related. But there is an interesting puzzle that one encounters in trying to specify exactly how they are related. It is very tempting to think that the ‘mental functioning’ sense of rationality is nothing but the fundamental sense, relativized to the beliefs of the agent, in place of the facts of the case. But this cannot be correct. For it may be that an action would be rational, in the fundamental sense, if the world were as my beliefs represent it, and yet it may still be that my performance of the action would be irrational in the ‘mental functioning’ sense. This may happen because I conspicuously lack a belief that I should definitely have: the belief that my action will cause me a great deal of suffering, for example. I may refuse to believe this, although I have more than enough evidence to believe it, because it may be that the suffering will be caused by someone I love, and I may deceive myself into thinking that the person would never hurt me. In such a case my action would be rational, in the fundamental sense, if my beliefs accurately represented the world. But it is nevertheless irrational in the ‘mental functioning’ sense. The next obvious strategy would be to define rationality in the ‘mental functioning’ sense in the following way: it is simply the same as the fundamental sense, but relativized to the beliefs that the agent should have, given the available evidence. But this strategy also fails, perhaps even more spectacularly. For it may be that an action would be rational, in the fundamental sense, if the world were as I should believe it to be, and yet it may still be that my performance of the action would

---


Conditions on an adequate theory

be irrational in the ‘mental functioning’ sense. How could this happen? It may be that, though I should believe that a certain unpleasant action will benefit me greatly in the long term, I do not actually believe it. In such a case, the fact that the action will benefit me (and that I should believe this) does nothing to mitigate the irrationality of performing it, if it would be irrational to do so in the absence of the future benefits. So two initially plausible accounts of the relation between the two senses of rationality are completely inadequate. And it is obvious that one cannot simply relativize to the set of beliefs that one does or should have, for this will typically be a set of inconsistent beliefs. Nor can one relativize to the beliefs one does and should have, for if one believes that a certain action will be quite painful, and will benefit no one, then it would be irrational to perform the action, even if one should not have this belief. This book provides an account of the relation between the ‘mental functioning’ and ‘fundamental’ senses of rationality in a way that not only avoids counterexamples, but also explains why the above relativizing definitions fail, and why they fail in the particular ways they do.

The ‘mental functioning’ and ‘fundamental’ senses of rationality are often distinguished by calling the former ‘subjective rationality’ and the latter ‘objective rationality,’ and this is the terminology I will use in this book. But quite often philosophers do not distinguish the two senses at all. And sometimes the fundamental sense is the only sense of rationality that is officially recognized, so that the phenomena captured by the ‘mental functioning’ sense end up being described with phrases such as ‘rational, relative to the beliefs of the agent.’ In earlier writing I sometimes borrowed a piece of terminology from Allan Gibbard, who uses the term “advisable” as a label for “[w]hat it makes sense to do objectively, in light of all the facts” – that is, for what I am calling ‘objectively rational’ action. Gibbard’s terminology has the advantage of minimizing the risk of thinking that the objective notion has much to do with mental functioning directly. However, I now prefer the terms ‘subjective rationality’ and ‘objective rationality’ because, despite the fact that a perfectly (subjectively) rational person might often perform objectively irrational actions, it is uncontroversial that there is a very close connection between subjective and objective rationality. Using the two terms ‘rationality’ and ‘advisability’ wrongly lends an air of plausibility to objections that depend

---

9 Gibbard (1990), p. 89.
on the false premise that a fully informed agent, performing a rational action, might nevertheless be performing an inadvisable one. Also, the ‘subjective/objective’ terminology allows me to use phrases such as ‘an account of rationality’ in order to indicate an account both of subjective and objective rationality, and of the relation between the two. In what follows, when I use the words ‘rational’ or ‘irrational’ without any qualification, they should be understood in the subjective sense, which fits more with the everyday understanding of these words.

Although I will not provide a full account of the relation between subjective and objective rationality until chapter 7, some limited claims about their relation are independently plausible, and will be very useful in a number of earlier arguments. First, if an agent knows all the facts relevant to his action, then if that action is objectively irrational – that is, if it is prohibited by the fundamental normative principle – it is also subjectively irrational. This connection will allow us to move from the objective irrationality of an action to its subjective irrationality (and therefore from its subjective rationality to its objective rationality) in all cases in which it is permissible to stipulate that the agent has all relevant beliefs. This claim is very similar to one of Gibbard’s: "in the special case in which I know all that bears on my choice, what is rational for me to do is what is advisable for me to do." My claim is slightly weaker, however, for it does not entail that we can always move from objective rationality – what Gibbard calls ‘advisability’ – to subjective rationality (or, therefore, from subjective irrationality to objective irrationality) even in the case in which the agent is fully informed. As chapters 7 and 8 will explain, this move can be illegitimate when the agent does not care about the considerations that make his action objectively rational, or only performs that action because of failures of instrumental rationality: cases in which an agent does ‘the right thing for the wrong reasons.’ These are cases in which the etiology of the action is what makes it subjectively irrational, and this gives rise to the possibility that if the same action had been done for other reasons it would have been subjectively rational – and therefore it also gives rise to the possibility that an action can be objectively rational despite being subjectively irrational, even in a fully informed agent. Acknowledging this possibility, we can make the following claim: if the agent is fully informed, then if his action would have been subjectively irrational no matter what its etiology, it is also objectively irrational. Since, as has already been noted, if

the action of a fully informed agent is objectively irrational then it is also subjectively irrational, Gibbard’s claim is very close to correct.\footnote{It may be worth mentioning at the outset that subjective rationality, as understood here, is not to be confused with Thomas Scanlon’s technical and restricted sense of rationality, according to which actions are rational or irrational depending solely on whether or not they are in line with the agent’s \textit{normative judgment} that he or she ought to perform the action. Scanlon’s sense is inadequate if we want a general notion that captures the wide range of failures of practical mental functioning that are relevant to questions of mental illness, competence to give consent, moral responsibility, and so on. One reason for its inadequacy as such a general notion is that, as I argue in chapter 9, our actions very rarely involve the normative judgments that are presupposed when one calls an action rational or irrational in Scanlon’s sense. This is not to criticize Scanlon’s choice of terminology. One is free to use whatever terminology one wants, as long as one is clear about what one means. It is only a reminder that Scanlon himself recognizes many other failures in practical mental functioning: insensitivity to certain reasons, compulsions, and phobias, even when they are accompanied by rationalizing normative judgments (or no normative judgment), etc. It is to this more general class that I am applying the term ‘subjectively irrational.’ Of course, acting against one’s considered judgment as to how one ought to act is our \textit{priors} of the sort of irrationality with which this book is concerned, and it will be captured by the account offered in chapter 7.}

Another interesting feature of subjective rationality is the following. It seems that if one is simply unmoved by awareness of the prospect of some significant harm for oneself – say, that one’s action will cause one a great deal of pain, or will risk some nontrivial injury – this does nothing to the normative force of the reason that one is aware of. This is not to deny that one can be perfectly rational in willingly suffering such harms, if there are sufficient countervailing reasons. But the fact that one needs significant countervailing reasons shows that a rational person cannot be very indifferent to such harms for himself. On the other hand, relative indifference to the harms that one’s actions may cause other people is not nearly as universally regarded as irrational in the ‘mental functioning’ sense that is relevant to questions of moral responsibility and so on. Rather, when we speak of such indifference, we use words such as ‘callous,’ ‘selfish,’ or ‘mean.’ No one denies that it is rationally permissible to be motivated by other-regarding reasons. But it does not seem to be rationally required to the same extent that it is rationally required that one avoid harms for oneself. Of course there are views of rationality according to which one is in fact required to be as strongly motivated by altruistic as by self-interested reasons. This introductory chapter is not the place to combat such views. Rather, it is the place to mention that such accounts need to make us comfortable with some apparently counterintuitive judgments as to whether certain actions are subjectively rational or not – rational in the sense that is relevant to questions of moral responsibility and so on. That
Brute Rationality

is, in order to succeed in convincing us that it is irrational, in this sense, to be indifferent to the harms one's actions will cause other people, they will have to account for the fact that we generally wish to hold extremely immoral people fully responsible for their sadistic actions.

But even if it is a mistake to defend the normative equivalence of self-interested and altruistic reasons, one cannot simply deny that there are such things as altruistic reasons. That is, even if it is only callous, and not irrational, to be indifferent to the harms one causes others, one should not therefore deny that it can be perfectly rational to make great sacrifices—even the ultimate sacrifice—for others. It would be a poor theory of rationality that insisted that it was irrational to sacrifice one’s life to save a group of strangers, or that held that one’s ‘real’ reason in such a case was essentially self-interested.12 If an agent is strongly motivated to save a group of other people, and acts accordingly at the cost of his own life, this may be perfectly selfless, and perfectly rational. These two facts—that it is rationally permissible to be selfish, but also rationally permissible to make selfless sacrifices for others—seem to suggest that whether or not one has an altruistic reason depends upon whether or not one has a corresponding altruistic desire. And indeed there have been philosophers who explicitly claim that while one’s objective interests or needs provide desire-independent reasons, there is another class of reasons, which includes altruistic reasons, that stem from one’s desires or values.13 The plausibility of such views derives entirely from their ability to capture some otherwise elusive phenomena. Moreover, such accounts will need some way of limiting the content of one’s reason-giving desires or values, so that they do not end up claiming that one has a reason to drink paint simply in virtue of a desire to do so, or that one has a reason to exterminate some offending race of human beings because of one’s racist values. It is one goal of the present book to provide an explanation for the differential impact of desire on the relevance of reasons to the subjective rationality of action. Moreover, this explanation will limit the importance of desires in such a way that one’s desire to drink paint or to hurt someone else never rationally justifies one’s action, while one’s desire to help someone else can provide such justification.

Because the following arguments will be so much at odds with the Kantian view that moral requirements are also rational requirements, it may seem as though they must be concerned with a more stringent notion of

12 It would be as poor as a theory that insisted that it would be irrational not to make such a sacrifice, because one would, by failing to act, cost more lives than one saved.