Jacob Ross

Rationality, Normativity, and Commitment[1]

Is rationality normative, in the sense that we ought to be rational in our actions and attitudes? Recently, the claim that rationality is normative has faced several challenges. In this paper, I will take up these challenges, and I will attempt to vindicate the normativity of rationality in the face of them. I will begin, in part 1, by clarifying the question at issue, and outlining the challenges to the normativity of rationality. Then, in parts 2 through 4, I will discuss and criticize some responses that have been offered to these challenges in the literature. And in the last two parts of the paper, I will offer my own unified response to these challenges.

1 Challenges to the Normativity of Rationality

Here I will begin, in section 1.1, with a partial clarification of the claim that rationality is normative. Then, in section 1.2, I will discuss three challenges to this claim, which I will call the *ignorance problem*, the *pragmatic reasons problem*, and the *mere incoherence problem*. In the following three sections I will examine some responses that have been given to these three challenges, respectively, and I will argue that each of these responses is inadequate.

1.1 The Question of the Normativity of Rationality

There are several alternative ways of understanding the problem of the normativity of rationality. To a first approximation, we may say that rationality is normative just in case one ought to be rational. Or, more precisely, we may say that rationality is normative just in case, for any agent, S, and any property, A, if rationality requires that S be A, then S ought to be A. But this first pass is not entirely satisfactory. For there may be some *ought*s that are not normative, or at least that are not always normative. Thus,

[1] In writing this paper, I benefited from feedback from audiences at Rutgers University and at the 2010 Metaethics Workshop in Madison, Wisconsin. Thanks to Jamie Dreier, Kenny Easwaran, Stephen Finlay, Shieva Kleinschmidt, Mark Schroeder, Sam Shpall, Julia Staffel, and Gideon Yaffe. Special thanks to Derek Parfit, for very detailed comments on every section of this paper.

one might hold that, where the *ought* under consideration is the *ought* of etiquette, one ought always to place one's dessert fork to the left of one's dinner fork. And yet there may be some contexts in which one has no reason whatsoever to act in this way.

This suggests an alternative conception of what it is for rationality to be normative. Perhaps to say that rationality is normative is to say that, for any agent S and property A, if rationality requires that S be A, then S *has some* reason to be A. But this is too weak. For if rationality is normative, then presumably one is somehow failing, or getting things wrong, if one violates the requirements of rationality. But one can do something one has reason not to do, or have an attitude that one has reason not to have, without failing in any way, or getting anything wrong—this would be true, for example, if one had stronger reason for some alternative action or attitude.

Thus, the claim that rationality is normative does not appear to be captured perfectly either by saying that we ought to be A when rationality requires us to be A, or by saying that we have reason to be A when rationality requires us to be A.[2] My preferred understanding of this claim is a precisification of the former account. To say that rationality is normative is to say that, for any subject S and property A, if rationality requires S to be A, then S ought *in a normative sense*, to be A. (I say "*a* normative sense" rather than "*the* normative sense" so as to allow for the possibility that there may be more than one normative sense of "ought"). This doesn't entirely clarify the problem, but only shifts the burden from one of understanding what it is for a property (rationality) to be normative, to one of understanding what it is for a concept (in particular, an *ought* concept) to be normative. This latter burden is one that I will take up later, in section 5.1.

1.2 The Three Challenges

One reason to doubt that rationality is normative arises from cases where agents are ignorant of some relevant facts, with the result that the actions or attitudes that would be most rational for them differ from those that are most favored by objective reasons. Consider Bernard Williams gin/petrol case. Suppose Genevieve is at a party, and she

[2] Derek Parfit has pointed out to me a third way of understanding the claim that rationality is normative: namely, as the claim that we *have reason to want to be*, *or to try to be*, rational. But once again, this seems too weak. For if we have stronger reason to want, or to try, *not* to be rational, then even if we don't so much as want, or try, to be rational, we needn't be failing in any way.

wants to drink a glass of gin and tonic, as she recognizes that, at the moment, there is nothing she would enjoy more. Her host, whom she has every reason to regard as trustworthy, prepares for her, before her very eyes, what he describes as, and what appears to be, a glass of gin and tonic. What he has prepared, however, is in fact a cleverly disguised glass of petrol. In this case, given that she has every reason to believe that the glass contains gin and tonic, and that she knows that there is nothing she would enjoy more than to drink a glass of gin and tonic, it seems the most rational course of action for her to take would be to drink from the glass. And yet, given that the glass actually contains petrol, it seems equally clear that, objectively speaking, or relative to all the relevant facts known or unknown to the agent, she ought not to drink from the glass. And so this seems like a clear case where, objectively speaking, an agent ought not to do what it would be most rational for her to do.

Note that this problem arises not only for actions but also for attitudes. Not only does rationality require that she drinks from the glass, it also requires her to *intend* to drink from the glass, and for her to *prefer* drinking from the glass to not drinking from the glass. And yet, given that the glass contains petrol, it seems she objectively ought not to intend, or to prefer, drinking from it. Thus, in cases involving ignorance, there can be a divorce between the actions and attitudes that it would be most rational to do or have, and the actions and attitudes that one ought objectively to do or have. Let us call this the *ignorance problem*.

A second reason to doubt the normativity of rationality arises in cases where one would be rewarded for having irrational attitudes, or punished for having rational ones. Suppose, for example, that Floyd believes, in spite of all the evidence to the contrary, that the Earth is flat. Flora, the president of the Flat Earth Society, has implanted a small explosive device into Floyd's brain, right next to his belief box. If he ever ceases to believe that the Earth is flat, this bomb will detonate, killing Floyd as a warning to other potential non-believers. In this case, it seems that Floyd has overwhelming reason to believe that the Earth is flat, namely, that if he doesn't his brain will explode. However, since the balance of evidence weighs strongly against the claim that the earth is flat, it seems he could not rationally believe this. Thus, it seems there can be cases where we ought to have irrational beliefs. And similarly for other attitudes. There could be a case

in which one will be punished terribly unless one intends to do something one has no reason to do, or that one cannot do, or cases in which one would be terribly punished for preferring an outcome that is clearly worse, or for fearing something that is clearly harmless. In any such case, it may appear that, in virtue of this overwhelming pragmatic reason, one ought to have an attitude that is irrational. Let us call this the *pragmatic reasons problem*.

As we will see later, the most common response to the pragmatic reasons problem is to draw a distinction between genuine reasons for (or against) attitudes and good- (or bad-) making features of attitudes that are not such reasons. However, once such a distinction is drawn, a third challenge arises for the normativity of rationality. For it seems that rationality can prohibit one from having a set of attitudes even when one has sufficient reason to have each of the attitudes in this set, considered on its own. For example, rationality may prohibit one from (intending to take only the bale of hay on the left, and intending to take the only bale of hay on the right) even when it seems one has sufficient reason for either of these intentions on its own. Similarly, rationality may prohibit one from (preferring chocolate ice cream to vanilla ice cream, preferring vanilla ice cream to strawberry ice cream, and preferring strawberry ice cream to chocolate ice cream) even when one has sufficient reason for any one of these preferences on its own. The problem, however, is that once we distinguish between genuine reasons and mere good- and bad-making features, it becomes doubtful that there can be any *reason* not to have these combinations of attitudes that rationality prohibits.

Such combinations of attitudes may, of course, have plenty of bad-making features, since having them may get one into trouble in various ways. But, it may be argued that a reason to ϕ must, at least in principle, be a reason *for which* one could ϕ. But it seems that we don't have, or fail to have, a combination of attitudes because of the good or bad-making features of this combinations of attitudes, but rather in virtue of considerations bearing on the *individual* attitudes of which it consists.[3] For example, we have or lack the belief that p because of evidence for or against p; we have or lack the intention to ϕ because of reasons for or against ϕ-ing; and we have or lack a preference for A over B

[3] See Appendix A of Parfit (2011).

because of the facts about how A and B compare. But if the only genuine reasons against a combination of attitudes are reasons against the constituent attitudes, then in cases where one has sufficient reason to have each of the constituent attitudes, it seems there cannot be any compelling reason not to have the combination of attitudes. And if we do not have compelling reason not to have a combination of attitudes, it seems it cannot be true that we ought not to have it. It seems, therefore, that in cases where we have sufficient reason for each of a set of incoherent attitudes, it is not the case that we ought to be coherent, and so it is not the case that we ought to be rational. We may call this the *mere incoherence problem*.

We will now consider, in turn, some responses that have been given to each of these problems, beginning with the ignorance problem.

2 The Ignorance Problem and the Objectivist Strategy

If we understand what an agent *ought objectively* to do, or *has most objective reason* to do, as what the agent ought to do relative to all the facts, including any facts of which the agent is unaware, then there is no denying that what an agent ought objectively to do can come apart from what would be most rational for her to do, as the gin/petrol case illustrates.[4] One response to this problem is to maintain that while what rationality requires can come apart from what we have most objective reason to do, there is still an essential connection between the two. Indeed, one might argue that what we are rationally required to do, or what we 'ought rationally' to do, can be defined in terms of the objective 'ought,' or in terms of objective reasons. And one might argue that the normativity, or apparent normativity, of the rational *ought* derives from this essential connection to the *ought* of objective reasons. Let us call this general approach to understanding the normativity of rationality the *objectivist strategy*.

Note that by the objectivist strategy, I specifically mean the strategy of understanding the rational 'ought' in terms of the *ought* of most objective reason, i.e., the *ought* that is relative to all the facts. For it is the latter ought whose dissociation from the rational ought gives rise to the ignorance problem. Thus, views on which the rational

[4] For important recent discussions of the alternative senses of "ought," see Parfit (2011) and Finlay (2009) and (2010).

ought is to be understood in terms of some other *ought* besides the objective one are not my present target.

2.1 Some Versions of Objectivism

Proponents of the objectivist strategy can adopt either of two views about the normativity of rationality. On the one hand, the objectivist might maintain that rationality has a genuine kind of normativity, in virtue of its connection with the objective *ought* or with objective reasons. Alternately, the objectivist might be an error theorist about the normativity of rationality, and claim that, in virtue of its connection with the objective *ought* or objective reasons, rationality *appears* to be normative, but this appearance is illusory.[5] Clearly, the objectivist must take the former line if she aims to vindicate the normativity of rationality. But here I will be concerned only with the general strategy of understanding rational requirements in terms of the *ought* of most objective reasons. The arguments I will provide will apply equally to any version of the objectivist strategy, regardless of whether it aims to vindicate, or to deflate, the normativity of rationality.

Several alternative proposals have been made as to how rational requirements are to be reduced to the objective *ought.* Tim Scanlon and Niko Kolodny have suggested that the actions and attitudes that are rationally required of us are those that we *believe* to be most supported by objective reasons.[6] Another view, suggested, at one point, by Ralph Wedgewood, is that the actions and attitudes that are rationally required of us are those that we *ought* to believe (given our evidence) to be most supported by objective reasons.[7] A third view, which is a simplification of the view offered in Parfit (2011), is that the actions or attitudes that are rationally required of us are those that *would* be most supported by objective reasons *if our actual descriptive beliefs were true*. That is, we are rationally required to A (where A is doing some action or having or lacking some attitude) just in case we have descriptive beliefs the truth of which would give us most

[5] This is the view argued for in Kolodny (2005).

[6] See Scanlon 1999 and Kolodny 2005.

[7] See Wedgwood 2003. I say "suggested", because Wedgwood does not fully endorse this view.

objective reason to A. Finally, there is a fourth view, proposed by Jonathan Way (2009), according to which rationality requires us to A just in case, given our evidence, we *ought* to have descriptive beliefs the truth of which would give us most objective reason to A. Thus, on the first two views just outlined, rationality is a function of one's actual or idealized beliefs about objective reasons, while on the third and fourth views just outlined, rationality is a function of the objective reasons that would obtain if our actual or idealized descriptive beliefs were true.

2.2 The Three Envelope Problem

Each of these four views has a number of difficulties, many of which I have discussed elsewhere.[8] For the present, however, I will focus on a single example that serves as a counterexample to all four of these views. I call it the *three envelope problem*.[9] Suppose that Chester must choose between three envelopes. He knows that the first envelope contains $900, and he knows that, of the two remaining envelopes, one contains $1000 while the other is empty. However, he has no idea whether the $1000 is in the second envelope or the third, and he has no evidence favoring either possibility over the other.

In this case, it seems clear that the rational thing for Chester to do is to take the first envelope. However, each of the four views just considered implies the opposite. For Chester knows that, either it is the case that he has most objective reason to take the second envelope, or else he has most objective reason to take the third envelope. Either way, it is not the case that he has most objective reason to take the first envelope. To the contrary, whether the $1000 is in the second envelope or in the third, Chester has most objective reason *not* to take the first envelope, since his doing so is incompatible with his doing what he has most objective reason to do. And Chester recognizes this. Hence, since he does not believe that he has most objective reason to take the first envelope, the Scanlon-Kolodny view implies that he is not rationally required to take the first envelope. And since he believes that he ought objectively *not* to take the first envelope, the Scanlon-Kolodny view implies that he is rationally required *not* to take the first envelope.

---

[8] Ross (2006).

[9] I introduced this case in Ross (2006). This case is structurally analogous to the mine-shaft case that was introduced in Donald Regan's *Utilitarianism and Cooperation*, and that Parfit discusses in *On What Matters*.

Furthermore, Chester has exactly the beliefs that he ought to have, given his evidence. And so the view Wedgwood suggests (according to which what we are rationally required to do is what we ought to believe that we ought objectively to do) has the same implications about this case as the Scanlon/Kolodny view: it implies that it is not the case that Chester is rationally required to take the first envelope, and it implies, to the contrary, that he is rationally required not to do so.

Now consider the third view, on which what we have most objective reason to do is whatever the truth of our descriptive beliefs would give us most objective reason to do. Since Chester's beliefs are in fact all true, and yet it is not the case that he has most objective reason to take the first envelope, it follows that Chester doesn't have any beliefs the truth of which would give him most objective reason to take the first envelope. And so the third view implies that it is not the case that Chester is rationally required to take the first envelope. Further, this theory implies that Chester is rationally required not to take the first envelope. For what explains the fact that Chester has most objective reason not to take the first envelope is that there is more money in either the second or the third envelope. And Chester *believes* that there is more money in either the second or the third envelope. And so he has beliefs the truth of which would give him most objective reason not to take the first envelope. Hence, the third view implies that he is rationally required not to do so. Further, since Chester has exactly the beliefs that he ought to have, given his evidence, Jonathan Way's theory (according to which what you are rationally required to do is what the truth of the beliefs you ought to have would give you most objective reason to do) has the same implications in this case as the third view.

2.3 Why Adopting Actualism Won't Solve the Three Envelope Problem

One might object to the above argument as follows:

> Your argument assumes *possibilism*. That is, it assumes that one *ought*, in the sense of having *most objective reason*, to ϕ just in case *doing the best that one could possibly do* would involve ϕ-ing. But the proponent of the objectivist strategy might instead adopt actualism. That is, she might adopt the view that one ought to ϕ just in case what one would actually do if one were to ϕ is better than what one would actually do if one were not to ϕ. And if she were to adopt actualism, then she could avoid the conclusion that Chester is rationally required not to take the first envelope. Consider, for example, the third view. On this view, Chester is rationally required not to take the first envelope just in case he has descriptive beliefs the truth of which

> would give him most objective reason not to take this envelope. Combined with actualism, this view implies that Chester is rationally required not to take the first envelope just in case he has descriptive beliefs the truth of which would make it the case that he has more objective reason to do what he'd actually do if he didn't take the first envelope than to do what he'd actually do if he did take the first envelope. But he has no such descriptive beliefs: it's perfectly consistent with his beliefs that, if he didn't take the first envelope, then he'd take the empty envelope. And so it's perfectly consistent with his beliefs that he has *less* objective reason to do what he'd actually do if he didn't take the first envelope than to do what he'd actually do if he did take this envelope. Similar remarks apply to the other versions of the objectivist strategy outlined above. And so the proponent of any of these versions of the objectivist strategy can get around the problem you have raised by adopting actualism.

There are two points worth noting in response to this objection. The first is that while actualism may enable the objectivist to avoid one of the counterintuitive implications we have observed (namely, that Chester ought not to take the first envelope), it does not enable the objectivist to avoid the other counterintuitive implication (namely, that it is not the case that Chester ought to take the first envelope). For, in the case under consideration (where Chester has precisely the descriptive and normative beliefs that he ought to have, given his evidence), all four versions of the objectivist strategy we have considered will imply that Chester ought to take the first envelope just in case he ought to believe that he has most objective reason to take the first envelope. But even on the actualist view, it is not true that Chester ought to believe that he has most objective reason to take the first envelope. For on the actualist view, Chester has most objective reason to take the first envelope just in case, were he not to take the first envelope, he would take the empty envelope. But in the case described, Chester lacks sufficient reason to believe that this conditional is true. And so, even assuming actualism, he lacks sufficient reason to believe that he has most objective reason to take the first envelope.

Thus, the adoption of actualism would not really solve the problem that the three envelope case poses for the objectivist. Moreover, when objectivism is combined with actualism, even worse problems arise, as can be seen from the following case:

> *Akratic Alcibiades*: Alcibiades is in critical condition, having seriously damaged his liver. So long as he protects his liver, he will live, but if he subjects it to any further stress, he will die. He has before him three glasses: a glass of water, a glass of rubbing alcohol, and a glass of molten wax, and he must drink the contents of exactly one of these glasses. If he drinks the water, his thirst will be relieved; if he

> drinks the rubbing alcohol, he will die; and if he drinks the molten wax, his mouth and esophagus will be burned, but he will not die. Because of his alcoholism, he will in fact drink the rubbing alcohol. In so doing, however, he will be acting against his better judgment, as he recognizes that he has more reason to choose either of his two alternatives.[10]

In this case, the actualist view implies that Alcibiades ought objectively to drink the molten wax, since his objective reasons favor what he'd actually do if he drank the molten wax over what he'd actually do if he didn't drink the molten wax (namely, drink the rubbing alcohol). Moreover, so long as Alcibiades recognizes that he is about to drink the rubbing alcohol, and so long as he recognizes what the outcomes would be of this three possible choices, actualism will imply that it follows from Alcibiades' descriptive beliefs that he ought objectively to drink the molten wax. And so the third version of objectivism, when combined with actualism, implies that Alcibiades is rationally required to drink the molten wax. This combination of view also implies that Alcibiades is rationally required to drink the water, since it follows from his descriptive beliefs that drinking the water is objectively better than what he'd do if he didn't drink the water. And so this view implies that Alcibiades is faced with a rational dilemma, since he is rationally required to carry out two mutually exclusive courses of action. But neither the claim that Alcibiades is rationally required to drink the molten wax, nor the claim that he is faced with a rational dilemma, is remotely plausible. And so the third version of objectivism has unacceptable consequences when combined with actualism. And if we stipulate that Alcibiades is fully rational in his descriptive beliefs and in his beliefs about objective reasons, and that Alcibiades recognizes the truth of actualism, the other versions of the objectivist strategy we have considered will have the same unacceptable implications.

2.4 The Generality of the Problem Facing Objectivism

I have considered four objectivist theories that attempt to understand what rationality requires of an agent in terms of her actual or idealized beliefs concerning, or bearing on, what she ought objectively to do. I have argued that each of these theories has

[10] I discuss cases of this kind in Ross (forthcoming).

unacceptable implications concerning the three envelope problem. One might suppose, however, that some other such reduction might be more successful. And so I will end this section with a more general argument that no such strategy can succeed. More generally, I will argue that what rationality requires of an agent cannot be understood in terms of the beliefs that the agent has or ought to have, and so, *a fortiori*, it cannot be understood in terms of the beliefs the agent has or ought to have *concerning, or bearing on, what she ought objectively to do*.

I will argue this by constructing two cases such that, between these cases, there is a difference in how the agent is rationally required to act, but no relevant difference in what the agent believes, or in what she ought to believe given her evidence. The basic strategy is to construct a pair of cases, Case 1 and Case 2, between which there is initially a difference in what the agent believes and ought to believe, but where, in both cases, the agent then learns some proposition, E, that is stronger than what he initially believed in either of these cases. Thus, after the agent has learned E, the two cases no longer differ with respect to what the agent believes or ought to believe. And yet these cases continue to differ with respect to the agent's actual credences, and with respect to what the agent's credences ought rationally to be, and, consequently, these cases continue to differ with respect to how the agent ought to act. Essential to my argument is the claim, forcefully defended in Williamson (2002), that phenomenal facts, such as facts about the intensity of pains, are non-luminous, so that there could be two pains belonging to the same agent, such that the first pain is more intense than the second, and yet the agent does not know, and would not be justified in believing, that the former pain is more intense than the second.

> Case 1: Arthur Dent has a tooth ache in each of his thirty-two teeth. He is then asked how many of his teeth hurt more than his upper right incisor. (Let x be the number in question.) As a matter of fact, $x = 16$. However, Arthur doesn't flat-out believe this. Initially, all he flat-out believes is that x is between 15 and 17. And initially, his credence is .25 that the $x = 15$, .5 that $x = 16$, and .25 that $x = 17$. And Arthur is fully rational in all these initial beliefs and credences. But then Arthur acquires some additional evidence. He learns from Mary, the perfectly reliable super-scientist with a brain-scan device, that x is either 16 or 17. After telling him this, Mary offers Arthur a bet on the proposition that $x = 16$. The bet costs \$1, and pays \$2 if the proposition is true.

Case 2: Here Arthur's initial situation, and Arthur's initial flat-out beliefs, are exactly as in Case 1, except for the following differences. In this case 17 (rather than 16) of Arthur's teeth hurt more than his upper right incisor. In other words, in this case $x = 17$. And what Arthur initially flat-out believes is that that x is between 16 and 18. Further, his credence is .25 that $x = 16$, .5 that $x = 17$, and .25 that $x = 18$. And Arthur is fully rational in all these initial beliefs and credences. As in Case 1, Arthur then learns from Mary that x is either 16 or 17, and Mary then offers Arthur the even-odds bet on the proposition that $x = 16$.

The only difference between Arthur's flat-out beliefs between these two cases is that in the former his strongest belief concerning the value of x is that it is between 15 and 17, whereas in the second case his strongest belief concerning x is that it is between 16 and 18. However, this difference is eliminated when Arthur learns from Mary that the x is either 16 or 17, for, after learning this, Arthur's strongest belief concerning x is the same in both cases, namely that it is 16 or 17. Thus, after learning this information from Mary, there is no longer any difference between the two cases in Arthur's flat-out beliefs.

There will remain, however, a difference between the two cases in Arthur's credences. Let p_{16} be the proposition that $x = 16$, and let p_{17} be the proposition that $x = 17$. In Case 1, Arthur's initial credence in p_{16} conditional on $(p_{16} \vee p_{17})$ is 2/3, whereas in Case 2, his initial credence in p_{16} conditional on $(p_{16} \vee p_{17})$ is 1/3. Consequently, upon learning $(p_{16} \vee p_{17})$, Arthur's unconditional credence in p_{16} will be 2/3 in Case 1, and 1/3 in Case 2. Thus, in Case 1, the rational thing for Arthur to do will be to accept the even-odds bet on p_{16}, whereas in Case 2 the rational thing for him to do will be to decline this bet. Thus, after Arthur learns from Mary that $(p_{16} \vee p_{17})$, while the two cases continue to differ with respect to Arthur's rational credences, the two cases no longer differ with respect to what Arthur (outright) believes, nor do they differ with respect to what it is rational for him to (outright) believe. And yet they differ with respect to what it is rational for him to do. Thus, what agents ought rationally to do does not supervene on their outright beliefs. And so, a fortiori, it does not supervene on their beliefs concerning, or bearing on, what they ought objectively to do. And so the objectivist attempt to reduce

questions about what agents are rationally required to do to questions about such beliefs cannot succeed.[11]

It is worth noting how little we need to assume in order to generate this problem for the objectivist strategy. All we need to assume is that there is some possible agent *x* and two possible worlds, w1 and w2, such that,

(i) *x* is fully rational in both w1 and w2.

(ii) In both w1 and w2, it is compatible with all of *x*'s beliefs that *x* is in w1, and it is likewise compatible with all of *x*'s beliefs that *x* is in w2.

(iii) In w1, *x*'s rational credence in w1 conditional on (w1 ∨ w2) is higher in w1 than it is in w2.

So long as these conditions obtain, it follows that if, in both worlds, *x* learns the disjunction (w1 ∨ w2), and rationally revises her beliefs in response to this information, then there will be no difference between these worlds in *x*'s subsequent beliefs: in both worlds she will come to believe just those propositions that follow from (w1 ∨ w2). And yet the two worlds will differ with respect to *x*'s posterior credences in w1 and in w2, and so they will differ with respect to how *x* ought rationally to act.

And if we assume that evidence is non-luminous, in the sense that agents are not always in a position to know what their evidence is, then it is very plausible that these conditions can obtain. Let w1 and w2 be two possible worlds that differ only in that, between them, there is a very slight difference in the evidential state that agent *x* is in (perhaps, for example, the upper right incisor of the agent hurts ever so slightly more in w1 than in w2). In this case, assuming anti-luminosity, it is plausible that, if *x* is fully rational in both worlds (and hence condition (i) obtains), w1 and w2 will each be compatible with all of *x*'s beliefs in both worlds (and hence condition (ii) will obtain). And, assuming that the fact that one is in a given evidential state provides some evidence for the proposition that one is in that evidential state, *x*'s rational credence in w1 conditional on (w1 ∨ w2) is higher in w1 than it is in w2 (and hence condition (iii) will obtain).

[11] In section 2.6, I will discuss the possibility that the objectivist might understand rational requirements in terms of credences or *degrees* of belief, rather than in terms of outright beliefs.

It seems, therefore, that the failure of the objectivist strategy follows from very minimal assumptions.

2.5 How this is a Problem for the View of Rationality as Reason-Responsiveness

One view of rationality, which is widely regarded as a truism, is that to be rational is to respond appropriately to reasons, or apparent reasons. The above argument, however, shows that this view, as it is often construed, is incorrect. For the truism is often understood to as implying the following:

> *Reasons Responsiveness*: For any agent x and action type ϕ, what rationality requires of x is a function of the reasons, or apparent reasons, that x has.

And the *reasons* (or apparent reasons) *that an agent has* are generally understood to be the set of relevant propositions to which the agent stands in some privileged epistemic relation, a relation which has been variously interpreted as *belief*, as *justified belief*, or as *knowledge*.[12] However, the toothache example shows that if the (apparent) reasons an agent has are understood in this way, then Reasons Responsiveness can't be right. For, after learning that x is either 16 or 17, Arthur believes exactly the same set of propositions in Case 1 as in Case 2. Similarly, assuming that, initially, the only difference in Arthur's justified beliefs and knowledge between the two cases is that in Case 1 he justifiedly believes and knows that x is between 15 and 17 and in Case 2 he justifiedly believes and knows that x is between 16 and 18, it follows that, upon learning that x is either 16 or 17, there will remain no difference between the two cases in his justified beliefs or in his knowledge. Thus, however one construes the privileged epistemic relation to a proposition that is constitutive of having a reason or apparent reason, the two cases will not differ with respect to the set of propositions to which Arthur stands in this privileged relation. Further, the two cases will not differ with respect to which, among these privileged propositions, are *relevant* in Arthur's situation, where he must choose between accepting and declining a bet on the proposition that $x =$

[12] For the distinction between there *being* a reason for s to phi and s *having* a reason to phi, see Williams, 1979. And for a discussion of the alternative conceptions the relation one must bear to a reason in order to count as *having* that reason, see Schroeder 2008 and forthcoming.

16. Consequently, if the (apparent) *reasons an agent has* are the relevant propositions to which this agent stands in the privileged epistemic relation, then the two cases will not differ with respect to the (apparent) reasons Arthur has. Consequently, if reason-responsiveness were true, then the two cases could not differ with respect to what is rationally required of Arthur.

It seems, however, that in Case 1, Arthur is rationally required to accept the even-odds bet on the proposition that $x = 16$, since his rational credence in this proposition is 2/3. By contrast in Case 2, he is rationally required not to accept this bet, since his rational credence in it is 1/3. This case seems to show, therefore, that we must either reject Reason Responsiveness, or else we must reject the standard conception of what it is to have a reason or apparent reason. But perhaps the objectivist needn't accept Reason Responsiveness. I will explore an alternative option in the next section.

2.6 Could a Generalization of the Objectivist Strategy Succeed?

Suppose the objectivist grants that what agents are rationally required to do cannot be understood in terms of their actual or idealized beliefs concerning, or bearing on, what they ought objectively to do. Even so, she needn't concede total defeat. For she might claim that there is some other way of understanding the normativity of rationality in terms of some objective normative notion. In particular, perhaps she can understand the normativity of rationality in terms of the normativity of objective *reasons* as represented by our *credences*.

To do so, she would need a way of quantifying the strength of the objective reasons favoring various possible outcomes, in the form of a measure of the objective value or objective choiceworthiness of outcomes for an agent at a time—what we may call an *objective value function*. (A traditional consequentialist will understand this objective value in agent-neutral and time-neutral terms, whereas a non-consequentialist will hold that the objective value of a given outcome can vary with respect to agents and times.) The objectivist could then understand rationality in terms of the *maximization of expected objective value*.[13]

[13] For views along these lines, see Oddie and Menzies 1992, as well as Wedgwood 2003.

This might allow the objectivist to understand the normativity of rationality in terms of the normativity of objective reasons. The basic idea is that rational agents aim to maximize objective value. Unfortunately, since a rational agent doesn't generally know the outcomes of her various options, she isn't generally in a position to decide to take whatever option has the greatest objective value, since she doesn't generally know which option fits that description. And so the best she can do, given her imperfect knowledge, is to choose the option with the highest expected value, thereby minimizing her expected shortfall from her primary aim of maximizing objective value. Thus, the rational requirement to choose the decision-theoretically rational option can be seen as deriving its normativity from our fundamental aim of maximizing objective value.

There are various ways in which one might try to spell out this basic idea, since the *expected objective values* might be understood either in terms of the agent's actual credences, or in terms of her idealized credences (perhaps, the credences that it would be most rational for her to have given her evidence), and, similarly, the expected objective values might be understood either in terms of the agents actual evaluation of outcomes, or in terms of her idealized evaluations. We needn't concern ourselves, however, with the differences among these views, but can instead consider them together by focusing on the cases where they converge. Consider, for example, the three envelope case, and assume not only that Chester has precisely the credences he ought to have given his evidence, but also that he assigns to each possible outcome the objective value he ought to assign to it. And assume, for simplicity, that the objective value of a given outcome for Chester is directly proportional to the amount of money he will receive if it obtains. In this case, regardless of which of the four above views one adopts, one should claim that the option of taking the first envelope has the highest expected objective value for Chester at the time of choice, since the expected values of his options of taking the first, second, and third envelopes can be represented as 900, 500, and 500, respectively.

But there is a problem. For how are we to understand the numerical values assigned by the objective value function?[14] In the above example, we simply assumed

[14] Note that the problem I am about to present doesn't require that the value of an outcome can be represented by a *single number*. For the problem applies equally to the view that the value of an outcome can be represented by a *range of numbers*. The problem does not apply to a view on which the values of outcomes can be represented only ordinally. But the objectivist who aims to understand the normativity of rationality in terms of the normativity of

that they line up with dollar values. But in non-monetary cases, calibrating values with dollars isn't generally an option. Suppose I'm deciding between going to the beach and staying home. I know it will be either sunny or rainy, and if I go to the beach and it's sunny I'll get a tan, whereas if I go to the beach and it's rainy I'll get wet. Suppose, further, that my rational credence that it will be rainy is .1, and that getting a tan has more objective value for me than staying home, and staying home has more objective value for me than getting wet. On the objectivist picture, I ought rationally to go to the beach just in case the difference between the objective values assigned to getting a tan and staying home is more than nine times the difference between the objective values assigned to staying home and getting wet. But how are we to understand these numbers?

The natural way to understand these numbers is in terms of choice under uncertainty. Thus, where A is more valuable than B which is more valuable than C, to say that the value-difference between A and B is *x* times the value-difference between B and C can be understood to mean that one ought to be indifferent between B and a gamble whose two possible outcomes are A and C and where the probability of C is *x* times the probability of A.

But now it seems that the objectivist is in trouble. For she is attempting to understand the *ought* of rationality in terms of objective value. But to do so, she needs to put values on a numerical scale. And the only way to make sense of this numerical scale is in terms of the choices an agent ought to make under uncertainty. But when we are interested in the choices an agent ought to make under uncertainty, the *ought* in question is not the objective *ought*—it is not the ought of what the agent has most reason to do, relative to all the facts. And so the proposal in question doesn't really succeed in understanding the normativity of rationality in terms of the normativity of the objective *ought*, since it explains the normativity of rationality in terms of a normative notion (the notion of what the agent ought to choose under uncertainty) which is non-objective. Thus, while it may perhaps be defensibly claimed that, when it comes to actions,

---

expected objective value can't adopt the latter view, for ordinal rankings don't suffice to generate expected values. Thus, if you want to maintain that, in the three envelope problem, the rational thing to do is to take the first envelope because that's the option with the highest expected value, then you must maintain that the values of the three outcomes can be represented in a cardinal scale.

rationality is *coextensive* with the maximization of expected objective value, there does not appear to be any possibility of a *reduction* of the former to the latter.[15]

The objectivist might attempt to resist this argument. She might claim that while it's true that objective values can only be understood in terms of choice situations with probabilistic outcomes, the probabilities may be understood as *objective chances*. The concept of an objective chance is presumably not itself a normative notion, and so, a fortiori, it is not a notion that involves non-objective normativity. It is therefore a notion that the objectivist can freely appeal to in offering a reductive account of the normativity of rationality. The objectivist might therefore propose the following: where the value of A exceeds the value of B which exceeds the value of C, to say that the value-difference between A and B is *x* times the value-difference between B and C is to say that one ought (in the sense of having most objective reason) to be indifferent between B and a gamble whose two possible outcomes are A and C and where the *objective chance* of C is *x* times the *objective chance* of A.

Unfortunately, such a maneuver cannot succeed. For the outcomes to which the objectivist will need to assign objective values may be outcomes that do not admit of objective chances other than zero or one. Suppose, for example, that Zipporah doesn't know whether there exists a necessarily existing being (NEB), but she knows that Moses knows, and that he will tell her if she asks him. Suppose, further, that there are three relevantly different outcomes: the best possible outcome for Zippora is that a NEB exists and she knows so, the second best outcome for Zippora is that she remain ignorant of whether a NEB exists, and the worst possible outcome for Zippora is that no NEB exists and she knows so. Suppose, finally, that Zipporah's rational credence that a NEB exists is .5. According to the view under consideration, Zipporah ought rationally to ask Moses whether a NEB exists just in case the expected objective value of her doing so exceeds the expected objective value of her not doing so. And, given her rational credences, this will true just in case the value of knowing that a NEB exists exceeds the disvalue of knowing that no NEB exists. But how is the objectivist to make sense of this value difference? On the view under consideration, for the value of knowing that a NEB exists

[15] The ideas in the above paragraph were inspired by a conversation I had with Jamie Dreier.

to exceed the disvalue of knowing that no NEB exists is for it to be the case that Zipporah ought to accept a gamble that has a .5 objective chance of resulting in her knowing that a NEB exists and a .5 objective chance of her knowing that no NEB exists. But the very idea of such a gamble is incoherent, for it requires that it be objectively chancy whether a NEB exists, which is impossible.

To sum up: if the objectivist is to adopt the strategy considered in section 2.6, then she will need some (more or less precise) way of assigning numerical values to the values of outcomes. But the only plausible way of doing so is in terms of the choices an agent *ought* to make in choice situations where, conditional on some of the available options, more than one relevant outcome has positive probability. These probabilities, however, cannot be understood in purely objective terms (e.g., as objective chances), but must instead be understood in more subjective terms (e.g, as the agent's credences, or as the credences that it would be most rational for the agent to have given her evidence). But the question of what an agent ought to do given her credences, or given the credences she ought rationally to have given her evidence, is not a question about what the agent ought objectively to do. Hence, the strategy under consideration fails to reduce the normativity of rationality to the normativity of the objective *ought*.

3 The Pragmatic Reasons Problem and the Two Kinds of Reasons Strategy

Let us now turn to the second challenge to the normativity of rationality. The problem, recall, is that there appear to be cases in which agents have overwhelming pragmatic reasons not to be rational, and hence in which it seems that agents *ought* not to be rational—e.g., the case in which Floyd will be killed unless he believes that the Earth is flat. In this section, I will consider a standard kind of strategy for solving this problem, and I will argue that it cannot succeed.

The general strategy is to distinguish between two kinds of reasons. For any attitude, there is what we may call the *right kind* of reasons, which are reasons of an appropriate kind to bear both on whether one ought to have this attitude. Then there is what we may call the *wrong kind* of reasons, which are reasons that do not bear on whether one ought to have this attitude. Instead, they bear only on whether one ought to *want* or to *strive* to have this attitude, and on whether it is rational to want or to strive to

have this attitude. And pragmatic reasons, such as the fact that one will be rewarded or punished for having, or for failing to have, a given attitude, are reasons of the wrong kind. Thus, the pragmatic reasons problem is illusory. In cases in which we seem to have overwhelming pragmatic reason to have irrational attitudes, what's really going on is that we have overwhelming reason to want, or to strive, to have such attitudes. Hence, while these may be cases in which we ought to want, or to strive, to be irrational, they are not cases in which we ought to be irrational. Let us call this strategy for responding to the pragmatic reasons problem the *two kinds of reasons* strategy.

Several alternate criteria have been proposed for distinguishing between the right kind of reasons and the wrong kind of reasons. According to one view, defended by Derek Parfit (2001) and Christian Piller (2001), the right kind of reasons for an attitude are *object-given* reasons (i.e., facts about the object of this attitude) whereas the wrong kind are *state-given* reasons (i.e., facts about the state of having this attitude). Thus, the fact that the truth of the proposition believed is entailed by one's evidence is a reason of the right kind for believing it, since it is fact about the proposition believed, whereas the fact one would be rewarded for believing the proposition is a reason of the wrong kind, since it is a fact about the state of believing it. According to another view, defended by Pamela Hieronymi (2005), the right kind of reasons for a given attitude are reasons that bear on the question the answering of which settles whether one has the attitude in question. Thus, the evidence for a proposition, *p*, is a reason for believing it, since it bears on the question *whether p*, which is the question the answering of which settles whether one believes *p*.

But there are cases that generate problems for the two kinds of reasons strategy, however the two kinds of reasons may be defined. Consider the following pair of cases.

> *The Original Kavka Case*: It is now noon on Monday. Twenty four hours hence, Gregory will have the option of drinking a toxin, which would result in his being ill for one day. If, tonight, Gregory has the intention to drink the toxin, an eccentric billionaire will put a million dollars in his bank account. He will get to keep this money regardless of whether he drinks the toxin tomorrow.[16]

> *The Buridan-Kavka Case*: It is now noon on Monday. Twenty four hours hence, Ascot will be faced with a choice between two identical bales of hay. If, at

[16] This case is based on the one in Kavka 1983.

> midnight tonight, Ascot intends to take the bale on the left then, regardless of which bale he takes on Tuesday, on Wednesday he will be tortured and killed. Currently (at noon on Monday) Ascot is aware of all these facts. He also knows that, five seconds hence, he will forget about the punishment for intending to take the bale on the left. Further, he knows that, whichever intention he forms, having forgotten about the punishment for intending to take the bale on the left, he will retain his existing intention and follow through with it. Lastly, he knows that he likes to leave trivial decisions until the moment of action, and so if he waits five seconds and allows himself to forget about the punishment associated with intending to take the bale on the left, then he will wait until tomorrow at noon to decide which bale of hay to take.

It is generally agreed that, in the original Kavka Case, it would be irrational for Gregory to intend to drink the toxin. Thus, if we to maintain that one ought never to be irrational, one must maintain that it is not the case that Gregory ought to intend to drink the toxin, even if it is the case that he ought to *want*, and to *try to bring it about*, that he intends to drink the toxin. The proponent of the two kinds of reasons strategy will explain this by claiming that the fact that Gregory will receive a million dollars if he intends to drink the toxin is a reason of the wrong kind to bear on what he should intend. And this is precisely what the standard accounts of the right-kind/wrong-kind of reasons distinction imply. Thus, according to the Parfit-Piller view, the fact that Gregory will receive a million dollars if he intends to drink the toxin is a reason of the wrong kind because it is a state-given reason (a fact about having the intention) rather than an object-given reason (a fact about the action intended). Similarly, on Hieronymi's view, the fact that he will receive a million dollars if he intends to drink the toxin is a reason of the wrong kind, because it does not bear on the question *what to do*, which is the question the answering of which would settle what he intends.

But now consider the Buridan-Kavka Case. Here it seems clear that it would be irrational for Ascot to intend to take the bale on the left. But the only reason against having this intention is that he would be punished for having it. And so now the proponent of the two kinds of reasons strategy is faced with a dilemma. Are rewards and punishments for having intentions reasons of the right kind, or of the wrong kind, for intentions? If they are reasons of the right kind, then it is difficult to avoid the conclusion that, in the original Kavka case, Gregory has most reason to intend to drink the toxin. And since it is generally agreed that such an intention would be irrational, on this view it

is hard to avoid the conclusion that Gregory ought to be irrational. Suppose, on the other hand, that rewards and punishments are reasons of the wrong kind for intentions. It follows, on the view we are now considering, that these reasons will not bear on what Ascot ought to intend in the Buridan-Kavka case. But, apart from these reasons, Ascot has sufficient reason to intend to take the bale on the left. If the proponent of the two kinds of reasons strategy adopts the second horn of the dilemma, she must claim that Ascot has sufficient reason to intend to take the bale on the left. Hence, if she grants, as seems obvious, that such an intention would be irrational, she must claim that Ascot has sufficient reason to be irrational.[17]

Perhaps there is a way out of this dilemma. Rather than maintaining that pragmatic reasons don't bear at all on what attitudes one ought to have, the two kinds of reasons theorist might maintain that they can have some bearing, but that they are lexically dominated by the right kind of reasons. Thus, pragmatic can never outweigh reasons of the right kind, but they can break ties between possibilities that are equally favored by reasons of the right kind. Because pragmatic reasons can't outweigh reasons of the right kind, the fact that Gregory will receive a million dollars for having an intention that the right kind of reasons oppose can't make it permissible, let alone obligatory, for him to have this intention. But because pragmatic reasons can break ties, such reasons can make it the case that Ascot has more reason to intend to take the bale on the right than to take the bale on the left.

[17] In correspondence, Parfit has told me that he never intended the view to apply to intentions. However, a similar problem can be raised even if the Parfit-Piller view is restricted to attitudes other than intentions. Consider preferences. It seems that, even in Buridan's ass cases, a rational agent who intends to take a given bale of hay will thereby prefer to take that bale of hay. If that's right, then in an ordinary Buridan's ass case, the agent could rationally prefer to take either bale of hay, since she could rationally prefer to take either bale. And so the proponent of the two kinds of reasons strategy must say that, in an ordinary Buridan case, the agent has sufficient reason, of the right kind, to *prefer* either bale of hay. Consequently, if she maintains that pragmatic reasons are not reasons of the right kind, then she must claim that, even in the Buridan-Kavka case (where pragmatic reasons are introduced), the agent has sufficient reason, of the right kind, to prefer to take the bale on the left, and hence that that the agent could rationally have this preference. It seems clear, however, that the agent could not rationally have this preference. And if she maintains that pragmatic reasons are reasons of the right kind, then she must claim that, in original toxin case, the agent could rationally prefer to drink the toxin. Hence the same dilemma that arises for intentions also arises for preferences.

This lexical priority view, however, runs into problems in cases where the outcomes of one's options depend on one's intentions, as in the following case:

> *Thinking Outside the Box*: Barbara Boxer knows that two minutes hence, she will have the opportunity to take either of two boxes, A and B. And she knows that, one minute hence, her brain will be scanned and her intentions will be registered. If, one minute hence, she has the intention to take box A, then $10 will be placed in each box, and $1,000,000 will be placed in her bank account simply for having the intention. If, on the other hand, she has the intention to take box B, then $10 will be placed in box A and $20 will be placed in box B, but nothing will be placed in her bank account. And if she does not have either intention, then no money will be placed in either box, nor will any money be placed in her bank account. As a matter of fact, Boxer intends to take box B. Thus, she will end up with a total of $20.

In this case, it seems clear that Boxer is irrational in intending to take box B, since she should instead have intended to take box A in order to get $1,000,010. Note that this case is very different from the original Kavka case. In the latter case, it would be irrational for Gregory to intend to drink the toxin, because he knows in advance that he will have insufficient reason to follow through with this intention at the time of action. But in Thinking Outside the Box, Boxer knows that, if she intends to take box A, then at the time of action both boxes will contain the same amount of money, and so she will have sufficient reason to follow through with this intention.

Since it appears that Boxer's only rational option is to intend to take box A, the proponent of the lexical view who wants to vindicate the claim that Boxer ought to be rational must claim that Boxer has at least as much non-pragmatic reason to intend to take box A as she has to intend to take box B. But this doesn't seem to be the case. If we exclude the pragmatic reason for intention (namely, that the intention to take box A will result in her receiving a million dollars) and focus only on the ordinary reasons given by the features of the actions under consideration, then it would seem that Boxer's reasons favor intending to take box B. After all, taking box B will increase her wealth by $20, whereas taking box A would increase it by only $10. And so the lexical priority view fails to predict that the intention to take A is more rational than the intention to take B.

Thus, the two kinds of reasons view seems to be in a quandary. If pragmatic reasons are taken to be the right kind of reasons, and to have comparable force to ordinary, non-pragmatic reasons, then this view will imply that in the toxin case, Gregory

ought to intend to take the toxin, and hence that he ought to be irrational. If, on the other hand, pragmatic reasons are taken to be the wrong kind of reason, and to have no force whatsoever, then this view will fail to imply that Ascot ought not to intend to take the bale on the left. And if pragmatic reasons are taken to have some force, but not as much as non-pragmatic reasons, then this view will fail to imply that Barbara ought not to intend to take box B.

Just as the objectivist strategy fails to provide an adequate solution to the ignorance problem, so the two kinds of reasons strategy fails to provide an adequate solution to pragmatic reasons problem. In the next section, we will consider a solution that has been proposed to the mere incoherence problem, and I will argue that it too is unsuccessful.

4 The Mere Incoherence Problem and the Individual Attitude Strategy

Let us now turn to the third challenge to the normativity of rationality, namely, the mere incoherence problem. Recall that this problem arises from cases in which an agent has a set of attitudes such that each one on its own seems fine, and yet together they are incoherent. The problem was that, since reasons appear to be reasons for or against particular attitudes, it's hard to see how it can be the case that one ought not to have incoherent sets of attitudes if one has sufficient reason for each of the constituent attitudes. One response to this problem is to say that putative cases of mere incoherence are illusory: whenever there is a genuinely incoherent set attitudes, there is always something wrong with at least one of the constituent attitudes. Hence, problems with sets of attitudes are never purely holistic: if there's a problem with the set of attitudes, it derives from one or more of the attitudes in it. On this view, coherence requirements on sets of attitudes can be explained in terms of requirements concerning individual attitudes. Let us call this the *individual attitude strategy*. Representatives of this strategy include Joseph Raz, Thomas Scanlon, and Niko Kolodny.[18]

The individual attitude strategy can take many forms, according to what is said to be wrong with the problematic attitude or attitudes within the incoherent set. One might hold that, when there is an incoherent set of attitudes, there is some attitude in this set that

[18] See Raz (2005), Scanlon (2007), and Kolodny (2007).

the agent has most objective reason not to have. Or one might hold that there is some attitude in this set such that the agent believes, or ought to believe, that she has most objective reason not to have it. To say either of these things would be to combine the individual attitude strategy with a kind of objectivism similar to that discussed in section 1.3. That is, it would be a strategy that explains the irrationality of sets of attitudes in terms of a problem with the constituent attitudes, and that explains the latter in terms of real or apparent objective reasons. And in practice, the representatives of the individual attitude strategy tend also to be objectivists, and so to adopt a position of this kind. However, the arguments of section 2 suffice to show that no such strategy can succeed. Consider a version of the three envelope case in which, as a matter of fact, the $1000 is in the second envelope, but the agent knows only that there is $900 in the first envelope and $1000 in either the second or the third envelope. And suppose that the agent, in spite of his ignorance, forms the intention to take the second envelope. In this case, his beliefs and intentions will be incoherent, in the sense of being jointly irrational. And yet he will have no attitude that he ought objectively not to have, nor will he have any attitude that he believes, or ought to believe, that he ought objectively not to have.

Thus, the most popular versions of the individual attitude strategy fail for reasons we have already discussed. But one can adopt the individual attitude strategy without combining it with objectivism. One might instead claim that whenever a set of attitudes is incoherent at least one of the constituent attitudes is irrational, without attempting to explain the irrationality of the individual attitude or attitudes in terms of objective reasons. Such a strategy would be immune to the arguments already given, and so requires a separate discussion.

Here's an illustration. Suppose Rainer believes both that it will rain tomorrow and that it will not rain tomorrow. Clearly, his beliefs are inconsistent. But to explain why Rainer should not have such inconsistent beliefs, we needn't assume that there is any fundamental and irreducible requirement of belief consistency that applies to sets of beliefs, and that Rainer goes wrong in virtue of violating this requirement. For, in this case, it seems he must violate a requirement that applies to individual beliefs. As Kolodny points out (2007, *p*. 233), it is plausible that one can rationally believe a given proposition only if one's evidence supports this proposition more than it supports its

negation. But in the case of two contradictory propositions, *p* and not-*p*, it clearly cannot be the case that one's evidence supports each of them more than its negation. And so on this view, if one believes both a proposition or its negation, then at least one of the beliefs must be insufficiently supported by the evidence, and so at least one of them must be individually irrational.

One problem with this explanation of the requirement not to believe both a proposition and its negation is that it does not generalize to sets of inconsistent beliefs involving more than two beliefs. It seems that what's going wrong with someone who believes that it's raining and believes that it's not raining is relevantly similar to what's going wrong with someone who believes that it's raining, and believes that it's snowing, and believes that it's not both raining and snowing. Consequently, we should expect the same explanation of what is going wrong to apply to both cases. But in the case of a set of three or more jointly inconsistent propositions, it may be true of each of the propositions in the set that the evidence gives more support to it than to its negation.

Nor can one solve this problem by adopting a more stringent requirement on evidence, requiring that, in order for it to be rational to believe a proposition, the evidential probability of this proposition must be above some threshold, θ, where $\theta > .5$. For if one holds that $\theta = 1$, then one will be committed to an implausible form of infallibilism according to which one can rationally believe a proposition only if one's evidence makes it absolutely certain. If, on the other hand, one holds that θ is less than 1, then there will be some cases in which one is committed to holding that it is rational to believe each proposition in a set of jointly inconsistent propositions, where such beliefs would appear to be irrational. This will be true, in particular, in lottery cases involving a sufficiently large number of tickets: here the view under consideration will imply that one can rationally believe, of each of the tickets in the lottery, that it will lose. But such a combination of beliefs would not appear to be rational.

Thus, the consistency requirement on beliefs poses a serious problem for the defender of the individual attitude strategy. While she can offer an explanation of the prohibition against believing both a proposition and its negation, the latter is simply a

special case of a more general requirement not to have inconsistent beliefs,[19] a requirement which she cannot easily explain.[20] Hence, when it comes to belief coherence, the individual attitude strategy is not very promising. I will conclude this section by arguing that it is equally unpromising in the practical sphere. This can be seen from the following case.

> ***Satan's Apple.*** Satan has cut a delicious apple into infinitely many pieces, labeled by the natural numbers. For each piece of the apple, Eve must choose whether to take it or decline it, and she must make all these choices simultaneously. If she takes merely finitely many of the pieces, then she suffers no penalty. But if she takes infinitely many of the pieces, then she is expelled from the Garden for her greed. Either way, she gets to eat whatever pieces she has taken. Eve is aware of all these facts. She has a very strong preference for remaining in the Garden over being expelled, and she has a mild preference for eating more of the apple over eating less of the apple.[21]

Now suppose that, for each piece of the apple, Eve intends to take it. These intentions, together with the beliefs and preferences described above, form an irrational set of attitudes. Thus, the proponent of the individual attitude strategy must claim that there is something wrong with at least one of Eve's attitudes taken on its own. And since there is clearly nothing wrong with her beliefs and preferences, there must be something wrong with at least one of her intentions. Hence, there must be at least one piece, x, of the apple such that Eve's intention to take piece x is irrational. Now if the intention to take piece x is irrational, then this must be so either in virtue of Eve's other intentions, or independently of Eve's other intentions. But it can't be the latter, since for any piece, x, of the apple, were it not for her intentions to take the other pieces of apple, she could rationally intend to take x. Nor can Eve's intention to take x be irrational in virtue of her

---

[19] In saying this, I do not mean to imply that it is always irrational to have inconsistent beliefs. The preface paradox suggests otherwise. The correct formulation of the consistency requirement will therefore either be restricted so as not to apply to such cases, or else defeasible so as to admit of exceptions. My point is only that the correct formulation of the consistency requirement will not apply merely to pairs of propositions.

[20] One might think that the objectivist could somehow offer an objectivist explanation of the coherence requirement on credences, and then use the latter requirement in order to explain the consistency requirement on outright beliefs. For a critique of this kind of approach, see Ross and Schroeder (forthcoming).

[21] This case is borrowed from Arntzenius, et al. (2004). They discuss two versions of this case, a diachronic version and a synchronic version, though here I am considering only the synchronic version. I discuss this case further in Ross (2010b).

intentions to take the other pieces of the apple. For, in virtue of her other intentions, Eve is guaranteed to be expelled from the Garden regardless of whether she takes *x*. And so, holding these other intentions fixed, the only difference between her intending to take *x* and her not intending to take *x* is that the former would result in her getting one more piece of the apple.

Thus, taken individually, none of Eve's attitudes is irrational. And yet, together they are clearly irrational. The irrationality of this set of attitudes cannot, therefore, be explained in terms of the irrationality of its constituents, but is rather fundamentally holistic. The case of Satan's Apple thus presents an insurmountable obstacle to the individual attitude strategy.

5 Groundwork for the Vindication of the Normativity of Rationality

The goal of the present part of the paper is to lay the groundwork for a unified solution to the three challenges to the normativity of rationality that we have considered.

5.1 The Deliberative Ought *as the Fundamentally Normative* Ought

The ignorance problem, or the problem of the dissociation between what rationality requires and what we have most objective reason to do, presents a genuine threat to the normativity of rationality only on the assumption that the *ought* of most objective reason has a greater claim to normativity than the *ought* of rationality. But this, I will now argue, is a mistake.

For an *ought* concept to be normative is for beliefs involving this concept to guide those who have these beliefs in an appropriate way. What is this way? It would seem that the strongest and most direct way in which an *ought* concept, O, could play such a guiding role is this: a fully rational agent who believes she stands in the relation, denoted by O, to ϕ, will thereby be motivated to ϕ. It would seem, therefore, that whatever *ought* concept plays this role should be regarded as fundamentally normative.[22]

Note, however, that the *ought* of most objective reasons does not satisfy this condition. For in the three envelope case, Chester believes that he has most objective

[22] Broome (unpublished) makes this point. See also Ross (2010a).

reason to do something other than taking the first envelope, and hence that he has most objective reason not to take the first envelope. But if Chester is fully rational, then this belief will not motivate him to refrain from taking first envelope. If, therefore, the fundamentally normative *ought* concept is the one that directly guides us by motivating us to conform with its requirements, then the objective *ought* is not the fundamentally normative *ought*.

There appears, however, to be another kind of *ought* that does play the required role. This is what we might call 'the *ought* of practical deliberation.' When we are genuinely deliberating (as opposed to, say, merely plumping, or merely ascertaining the necessary means to our ends), we weigh reasons for and against our alternatives in order to figure out what we ought to do, in some sense of 'ought.' Thus, in genuine deliberation, we are guided, at least implicitly, by the question "what sould I do?" or "what ought I to do?" And we ask this question not simply in order to satisfy our curiosity, but in order to make up our minds about what to do, that is, in order to form an intention. Thus, the role of the *ought* of practical deliberation is to guide our intentions, and thereby to guide our actions. And so the *ought* of deliberation plays the role that is constitutive of being fundamentally normative.

Consider one example of deliberation, the deliberation in which we can imagine Chester engaging in the Three Envelope case:

> I know there's $900 in the first envelope. If I take either of the other two envelopes, I might end up with $100 more, but I'd be just as likely to end up with nothing. Thus, neither of the other envelopes is worth the risk. Hence, I ought to take the first envelope. So that's what I'll do.

Clearly the *ought* that figures in this deliberation, and with which Chester is concerned when he asks the deliberative question 'what ought I to do?', is not the *ought* of most objective reason, but rather the *ought* of what makes most sense relative to the agent's state of information. We may conclude, therefore, that the fundamentally normative *ought* is the *ought* that is relativized to the agent's information state. (Here I will remain neutral as to how exactly the agent's information state should be characterized, but a natural way of thinking about the agent's information state is as a probability function. This probability function needn't coincide with the agent's actual credences, but may

instead coincide with the credences that it would make most sense for the agent to have given her perceptual state, memories, etc.)

5.2 Rational Commitment

The fundamental normativity of the deliberative *ought* will prove to be essential to the vindication of the normativity of rationality. But it won't suffice. For the belief that one ought, in the deliberative sense, to ϕ motivates one to ϕ by motivating one to intend to ϕ. And the objects of intention are ways of *acting*. Thus, it is only where ϕ-ing is a way of acting that the belief that one ought to ϕ will motivate one to ϕ, insofar as one is rational. In the case of attitudes, we do not form them by reflecting on what attitudes we ought to have, but rather by reflecting on their objects. For example, we form the belief that *p* by reflecting on *whether p*, not by reflecting on whether we ought to believe that *p*. And we form the intention to ϕ by reflecting on whether we ought to ϕ, not by reflecting on whether we ought to intend to ϕ. Similarly, we form the preference for A over B by reflecting on the relative merits of A and B, not by reflecting on what preference we ought to have. Thus, while the belief that one ought to ϕ (where ϕ is a way of acting) will motivate a rational agent to ϕ, the belief that one ought to have some attitude will not similarly motivate a rational agent to have this attitude.

Consequently, while the fundamental normativity of the deliberative *ought* may help us to understand why we ought to act rationally (as I will argue in section 6.1), it will not suffice to explain why we ought to have rational attitudes and combinations of attitudes. To explain the latter, we will need some way of connecting the ways in which one ought to act with the attitudes one ought to have.

The key to drawing such a connection is that all attitudes, or at least all attitudes that are rationally evaluable, are connected in some way to of actions. In particular, there is a sense in which all such attitudes seem to *commit* us to acting in certain ways under certain circumstances. This is clearest in the case of intentions. It is natural to say that the intention to ϕ commits us to ϕ-ing. But the language of commitment seems to apply to other mental states as well. Thus, there appears to be a sense in which the preference for apples over oranges commits one to choosing apples when faced with a choice between apples and oranges. And there likewise appears to be a sense in which a

credence of .5 in the proposition that Stewball will win the race commits one to buying a bet on this proposition if the utility of winning the bet exceeds the disutility of losing the bet, and to declining such a bet if the reverse obtains.

Clearly, the notion of 'commitment' at issue here is not the moral sense in which promising to ϕ commits one to ϕ-ing. How, then, are we to understand this relation of 'commitment' in which mental states stand to ways of acting? I suggest the following: a mental state commits one to acting in some way, it obviates the need to deliberate about whether to act in this way. More precisely, if mental state M commits S to ϕ-ing in C, then M serves as a surrogate for deliberation concluding in the decision to ϕ given C. Having formed mental state M, she no longer needs to weigh the reasons for and against ϕ-ing in C. For these considerations have already come into play in forming mental state M. Since these considerations have already settled the matter that S is to be in M, they have already, as it were, settled the matter that S is to ϕ in C, eliminating the need for any further deliberation.

This characterization of commitments seems to capture what's going on in the cases of commitment considered above. Once one has formed the *intention* to ϕ in C, one no longer needs to deliberate about whether to ϕ in C. Hence one can form the prior intention to ϕ in C at a time when one has opportunity to think carefully about the matter, thereby avoiding the need to deliberate about whether to ϕ in C at the time of action when one may no longer have the luxury of engaging in such deliberation. This, indeed, can plausibly be regarded as one of the primary functions of intentions. But other mental states play an analogous role. Once one has formed the preference for apples over oranges, one no longer needs to deliberate when faced with a choice between an apple and an orange. One can form the preference for apples over oranges when one has the opportunity of weighing the pros and cons of each (taste, nutritional value, convenience, etc.), thereby avoiding the need to weigh these considerations each time one is faced with a choice between an apple and an orange. Similarly, once one has formed a credence of .5 in the proposition that Stewball will win the race, one no longer needs to deliberate about whether to accept a given bet on his winning. One can form this credence at a time when one can weigh all the considerations relevant to whether Stewball will, and then simply rely on it when one is offered such a bet.

I will now argue for two principles connecting reasons for attitudes and reasons for the actions to which one is committed by these attitudes. These two principles will play an important part in the vindication of the normativity of rationality that I will offer in part 6.

5.3 The Commitment Transmission Principle

If committing mental states are to play the role just described, as surrogates for deliberation, then this imposes restrictions on what can be a sufficient reason for such a mental state. Suppose that for some mental state, M, some action type, ϕ, and some circumstance, C, one could have sufficient reason (relative to one's evidence) to have M without having sufficient reason (relative to one's evidence) to ϕ given C. In this case, M would not serve as a surrogate for deliberation concluding in the decision to ϕ in C. For in this case, the fact that one is in mental state M, and that one is in this state for good reason, would leave it open whether to ϕ in C, and hence it would not eliminate the need to deliberate about whether to ϕ in C. Hence, by the above criterion of commitment, M would not commit one to ϕ-ing in C. Thus, if one can have sufficient evidence-relative reason to have M without having sufficient evidence-relative reason to ϕ in C, then M does not commit one to ϕ-ing in C. An equivalent way of stating this claim is as follows:

> *Commitment Transmission Principle*: if M commits one to ϕ-ing in C, then one has sufficient evidence-relative reason to have M only if one has sufficient evidence-relative reason to ϕ in C.

(Henceforth, I will drop the phrase "evidence relative." Unless otherwise stated, by "ought," "sufficient reason," etc., I will mean "ought relative to one's evidence," "sufficient reason relative to one's evidence," etc.) Now for any option, ϕ, to say that one has sufficient reason to ϕ in C is to say that it is not the case that one ought, or has most reason, not to ϕ in C. And so another way to state the Commitment Transmission Principle is as follows: if M commits one to ϕ-ing in C, then if one ought not to ϕ in C, then one ought not to have M. The Commitment Transmission Principle thus implies that compelling reasons *against* are transmitted from ways of acting to mental states that commit one to those ways of acting.

5.4 The Commitment Agglomeration Principle

Before attempting to respond to the challenges to the normativity of rationality, we will need one further principle. The Commitment Transmission Principle tells us that we have sufficient reason to have mental state M only if we have sufficient reason to act in the ways to which M commits us to acting. In order to apply this principle, we need to be able to figure out how we are committed to acting in virtue of having a given mental state.

In the case where M is some individual attitude, we can answer this question by appealing to the criterion for commitment proposed in section 5.2: attitude A commits one to ϕ-ing in C just in case attitude A functions, inter alia, as a surrogate for deliberation concluding in the decision to ϕ in C, and thereby settles the issue as to whether to ϕ in C. But what if M is a complex mental state involving a plurality of attitudes? In that case what will M commit one to doing?

Suppose, for example, that M involves a pair of attitudes, A_1 and A_2, and that A_1 commits one to ϕ-ing in circumstance C_1, and A_2 commits one to ψ-ing in circumstance C_2. In this case, M will involve one attitude that settles, in the affirmative, the question as to whether to ϕ in C_1, and another attitude that settles, in the affirmative, the question of whether to ψ in C_2. And so being in mental state M will itself settle, in the affirmative, each of these questions, and thus it will stand in both for deliberation concluding in the decision to ϕ in C_1, and for deliberation concluding in the decision to ψ in C_2. But if M settles in the affirmative whether to ϕ in C_1, and it likewise settles in the affirmative whether to ψ in C_2, then it settles in the affirmative whether to *ϕ in C_1 and ψ in C_2*. And so it stands in for deliberation concluding in the decision to act in both of these ways. Thus, by our criterion of commitment, it will commit one to *ϕ-ing in C_1 and ψ-ing in C_2*.[23]

[23] Mike Titelbaum objects that the case of Satan's Apple constitutes a counterexample to the commitment agglomeration principle. It would constitute such a counterexample if the following claims were both true, where C represents the choice situation in Satan's apple (i) for each piece α of the apple, Eve's preference for α commits her to taking α in C; (ii) the mental state consisting in the combination of these preferences does not commit Eve to taking all the slices of the apple. However, (i) is false. For in the case of Satan's Apple, for any given piece of the apple, the fact that Eve prefers receiving it to not receiving it does not settle the issue as to whether to take that piece of the apple.

We have considered a simple case involving a mental state that consists in having a pair of attitudes. But same reasoning will apply to mental states involving arbitrarily many attitudes. And so we can generalize, as follows:

> *Commitment Agglomeration Principle*: If a mental state M consists in having some set of attitudes, then M commits one to acting in the conjunction of the ways in which one is committed to acting by having the mental states in this set.

So much for laying the groundwork. It remains to respond to the three challenges to the normativity of rationality.

6 Responding to the Challenges

Earlier we considered three challenges to the normativity of rationality. The ignorance problem challenges, inter alia, the claim that we ought never to act irrationally; the pragmatic reasons problem challenges the claim that we ought never to have irrational attitudes, and the mere incoherence problem challenges the claim that we ought never to have irrational combinations of attitudes. In this concluding part of the paper, I will take up each of these challenges, and argue that we ought never to be irrational in any of these three respects.

6.1 Why We Ought Not to Act Irrationally

In section 5.1, I argued that the ignorance problem rests on a mistake. Recall that the ignorance problem arises from cases in which what an agent ought rationally to do comes apart from what she ought objectively to do, that is, from what she has most reason to do relative to all the facts. This will pose a threat to the normativity of rationality only on the assumption that objective *ought* is the fundamentally normative *ought*. And this assumption, I argued in section 5.1, is false.

I will now argue, positively, that what an agent ought rationally to do cannot come apart from what she ought, in the fundamentally normative sense to do. Recall that I argued, in section 5.1, that the *ought* concept that is fundamentally normative or action-guiding is the *ought* of practical deliberation. And the latter, I argued, is the *ought* that is relativized to the agent's information state. That is to say, what an agent ought, in the

deliberative sense, to do is whatever it makes most sense for the agent to do relative to her information state.

But surely *what it makes most sense for the agent to do relative to her information state* is the same as *what it would be most rational for the agent to do*. It seems, therefore, that we are in a position to make the following chain of identifications. The fundamentally normative *ought* is the *ought* of deliberation. What an agent ought to do, in the deliberative sense, is whatever it would make most sense for her to do relative to her information state. And what it would make most sense for her to do relative to her information state is whatever it would be most rational for her to do. Therefore, what an agent ought to do, in the fundamentally normative sense, coincides with what it would be most rational for her to do. Thus, once we recognize the fundamentally normative character of the *ought* of deliberation, for which I argued in section 5.1, the problem of the normativity of rationality, at least in relation to actions, dissolves.

Here one might raise the following objection.

> By employing the vague notion of an information state, you have obscured an important difference between an agent's evidence and an agent's beliefs. What an agent ought to do in the deliberative sense is whatever makes most sense relative to her *evidence*, whereas what she ought rationally do is whatever makes most sense relative to her *beliefs*. Consider, therefore, a case where an agent's beliefs are irrational. Consider, in particular, a modified version of gin/petrol case where all the evidence suggests that the glass contains petrol (which Bernard aims to avoid drinking), and yet Bernard irrationally believes that the glass contains gin (which he aims to drink). In this case, what Bernard ought *rationally* to do is to drink from the glass. But what he ought to do in the *deliberative* sense is to refrain from drinking from the glass. Hence, even if you are right that the fundamentally normative *ought* is the deliberative *ought* that is relative to the agent's evidence, it will still be true, in the present case, that Bernard ought, in the fundamentally normative sense, not to do what rationality requires. And so your attempt to vindicate the normativity of rationality fails.

This objection turns on the claim that what an agent ought rationally to do is whatever makes most sense in relation to her *actual* beliefs. And while a surprising number of philosophers appear to accept this claim, it is contrary to the ordinary, pretheoretic understanding of rationality. Suppose a safe has been dropped from the roof a sky scraper, and that it is plummeting toward the head of Wiley. Suppose further, that Wiley is looking up and sees the safe clearly and distinctly, and that he wants above all to avoid being hit by the safe. Suppose, further, that all of Wiley's evidence clearly indicates that

stepping out of the way is a necessary means to avoiding being hit by the safe. In this case, common sense says that Wiley's only rational option is to step out of the way of the safe. But according to the view of rationality we are now considering, the case is underdescribed: from the description of the case, nothing whatsoever follows about how it would be rational for Wiley to act. For that depends on what he believes: if Wiley happens to form the belief that stepping out of the way is a necessary means to avoiding being hit by the safe, then of course he ought rationally to do so. But if he doesn't happen to form this belief, then it would be perfectly rational for him to stay put and twiddle his thumbs as he watches that safe descending upon him.

Thus, it appears that the conception of rationality that underlies the present objection is highly counterintuitive. But given the popularity of this view, it may be worthwhile briefly considering what reasons one might have for accepting it. One possible reason is the following:

> Rationality is a matter of having mental states that are related *to one another* in the right kind of way, not a matter of having mental states that are related in the right way to things outside one's head. But while beliefs are mental states, evidence consists in facts that are outside the head. And so, while practical rationality can depend on the former, it can't depend on the latter.

This justification for the belief-based view of practical rationality is not very compelling. For even if evidence is understood in terms of mind-independent facts, the state of *possessing* evidence is itself a mental state. Consequently, the view that what is practically rational for an agent depends on the evidence she possesses is perfectly compatible with the claim that rationality is to be understood in terms of relations among mental states. And while there may be some relevant evidence that an agent does not possess, it's far from obvious that such evidence bears on what the agent ought, in the deliberative sense, to do.

Moreover, it cannot be plausibly maintained that rationality is simply a matter of interrelations among beliefs, intentions, preferences, and the like, and that it is not a matter of how such attitudes are related to the mental states the constitute possessing evidence (such as perceptual states and memory states). For it is almost universally granted that what it is rational for an agent to *believe* can depend on her perceptual states, memory states, etc. Thus, if rationality consists in proper relations among relevant

mental states, then perceptual states, memory states, etc., must figure among the relevant mental states. What reason could there be, then, for denying that such states could be relevant to practical rationality?

Let me consider one further possible justification for the belief-based view of practical rationality, which can be stated as follows:

> Everyone grants that one central kind of rational requirement is the one expressed by the instrumental or means-end principle. The details of this principle are controversial, but to a first approximation, this principle prohibits failing to intend to ψ when one intends to ϕ and believes that ψ-ing is a necessary means to ϕ-ing. But the correct formulations of the principles of rationality all have narrow scope: they state that, *if* one has certain attitudes, *then one must* have (or fail to have) some other attitude. Hence, the instrumental principle must say something like this: if you intend to ϕ, and you believe that ψ-ing is a necessary means to ϕ -ing, then you are rationally required to intend to ψ. But if this is the correct formulation of the instrumental requirement, then what intentions are rational will depend on an agent's actual beliefs. Consequently, what it is rational for an agent to do will likewise depend on her actual beliefs.

This objection rests on the claim that the proper formulations of the requirements of rationality all have narrow-scope, in the sense that such formulations state that, if an agent has certain attitudes, then she must have, or fail to have, some other attitude. However, the case of Satan's Apple discussed in part 4 shows that the narrow-scope conception of rational requirements is inadequate. For in the case of Satan's Apple, there are certain combinations of attitudes that rationality requires Eve not to have. In particular, it prohibits her from being such that, for every piece α belonging to some infinite subset of the pieces of the apple, she intends to take α. But this is not a narrow scope requirement, nor can it be derived from any narrow scope requirement: it cannot be derived from requirements of the form (if you have combination of attitudes M, then you must have (or fail to have) attitude A). Thus, the case of Satan's apple demonstrates that we must grant that some requirements of rationality require a wide-scope formulation. And if we grant this, then why not grant that the instrumental principle itself requires a wide-scope formulation? Perhaps the instrumental principle should be stated as follows:

> *Means-End Coherence*: Rationality requires that one not (intend to phi, believe that ψ -ing is a necessary means to phi-ing, and fail to intend to ψ).

And if we adopt this wide-scope conception of the instrumental requirement, then it will no longer follow from this requirement that what it is rational for an agent to do depends on her actual beliefs. For on the wide scope conception, someone who actually intends to ϕ, believes that ψ-ing is a necessary means to ϕ-ing, and fails to intend to ψ, can come into compliance with the instrumental principle either by forming the intention to ψ, or by dropping the intention to ϕ, or by dropping the belief that ψ-ing is a necessary means to ϕ-ing.

To sum up: we can solve the ignorance problem by recognizing that what an agent ought, in the fundamentally normative sense, to do is whatever makes most sense relative to her evidence, and that this coincides with what it would be most rational for her to do. The main grounds for rejecting this identification is the view that what an agent is rationally required to do depends not on her evidence, nor on the beliefs that would be most rational given her evidence, but instead on her actual beliefs. I have argued, however, that this view is inherently implausible, that one possible motivation for this view rests on a mistake, and that another of its possible motivations is undermined by the case of Satan's Apple.

6.2 Why We Ought Not to Have Irrational Attitudes

I have argued that we ought, in the fundamentally normative sense, not to *act* irrationally. It remains to be shown, however, that we ought likewise not to have irrational *attitudes*. The key to showing this is the Commitment Transmission Principle, for which I argued in section 5.3. For it follows from this principle that *reasons against* are transmitted from ways of acting to attitudes that commit one to acting in those ways. Thus, if we ought normatively not to act in certain ways, then we ought normatively not to have attitudes that commit us to acting in these ways (though it may of course be true that we ought to *cause* ourselves to have such attitudes, as in the cases of "rational irrationality" discussed in Parfit (1984)). Therefore, given the Commitment Tranmission Principle, in order to show that an agent ought normatively not to have a certain attitude, it suffices to show that her having this attitude would commit her to acting in ways in which she ought, normatively, not to act. And hence, given what I argued in section 5.1, it suffices to show that her having this attitude would commit her to acting in ways in

which she ought, *in the deliberative sense*, not to act—i.e., that this attitude would commit her to acting in ways in which she ought not to act relative to her evidence. And so this is what I must now show.

I will now argue that irrational attitudes commit those who have them to doing things they ought not to do—where *ought* is here understood in the deliberative sense, the sense which we earlier argued is fundamentally normative. Hence, it follows from the Commitment Transmission Principle that we ought, in this robustly normative sense of *ought*, not to have irrational attitudes. This will be so even if we would be greatly rewarded for having such attitudes. For the fact that one will be greatly rewarded for having an irrational attitude does not change the fact that having it would commit one to acting as one ought not to act. Thus, the Commitment Transmission Principle solves the pragmatic reasons problem, and succeeds, I will argue, where the two kinds of reasons strategy fails.

First consider belief. Recall the case of Floyd, who will be killed unless he believes that the Earth is flat. Now it is irrational for one to believe that *p* just in case one lacks sufficient evidence for *p*. Thus, since Floyd has much evidence that the Earth is round, and scant evidence that it is flat, it is irrational for him to believe that the Earth is flat. Thus, to solve the pragmatic reasons problem, we must maintain that Floyd ought not to believe that the Earth is flat, in spite of what may appear to be an overwhelming pragmatic reason for doing so. And the Commitment Transmission Principle implies precisely this. For believing the Earth is flat would commit Floyd to *acting as if* the Earth is flat, for example, by accepting bets on the proposition that the Earth is flat. But Floyd does not have sufficient reason to act in such ways. Since believing the Earth is flat would commit him to acting in ways in which he ought not to act, it follows from the Commitment Transmission Principle that he ought not to have this irrational belief.

Next consider preferences. Suppose I'll be killed unless I prefer that I undergo a severe pain on Tuesday to a slight discomfort on Wednesday. Such a preference would nonetheless be irrational. And it follows from the Commitment Transmission Principle that I ought not to have this preference. For it would commit me to *choosing* severe pain on Tuesday over slight discomfort on Wednesday, and that's something I ought not to do. In general, if one lacks sufficient reason to choose A over B, then one will lack sufficient

reason to something one is committed to doing by the preference for A over B, namely choosing A over B when presented with a choice between these alternatives. Hence, it follows from the Transmission Commitment Principle that one will lack sufficient reason to prefer A to B.

Now consider intentions. The Commitment Transmission Principle can easily explain why rewards for having intentions do not provide sufficient reason for them. Hence, it can explain why, in the original Kavka case, Gregory does not have sufficient reason to intend to drink the toxin, despite the fact that he would be rewarded for having this intention: for this intention would commit him to doing something he lacks sufficient reason to do, namely, drinking the toxin. What about the Buridan-Kavka case, where Ascot would be punished for intending to take the bale on the left? Here, it seems it would be irrational for Ascot to *intend* to take the bale in the left, despite the fact that he has sufficient reason to *take* the bale on the left. Hence, if we are to maintain that Ascot ought not to have any irrational attitudes, we must maintain that he ought not to intend to take the bale on the left. One might think, however, that the Commitment Transmission Principle could not explain this. For one might think that the only thing intending to take the bale on the left commits Ascot to doing is taking the bale on the left, which is something he has sufficient reason to do.

This, however, would be a mistake. Intending to ϕ doesn't just commit one to ϕ-ing, for it doesn't just settle the question as to whether to ϕ. It also settles the question as to whether to deliberate further about whether to ϕ in the absence of new evidence coming to light. And so it serves as a surrogate not only for first-order deliberation concerning whether to ϕ, but also for second-order deliberation concerning whether to engage in further deliberation about whether to ϕ. Thus, in intending to ϕ, we commit ourselves to not reopening the question as to whether to ϕ unless new relevant information comes to light. But suppose that on Monday Ascot were to intend to take the bale on the left on Tuesday. In this case, unless he changes his mind, he will retain this intention until midnight, and as a result he will be tortured and killed on Wednesday. Now if he reopens deliberation, he may change his mind, but if he does not reopen deliberation, he will not change his mind. Thus, if Ascot were to intend to take the bale on the left, then it would be the case that he ought to reopen deliberation.

Now let C be a circumstance in which it is not yet midnight on Monday night, and in which Ascot intends to take the bale on the left. As we have just seen, Ascot has most reason to reopen deliberation in C. But intending not to drink the toxin would commit Ascot to not reopening deliberation in C. And hence, the intention to take the bale on the left commits Ascot to acting in C in a manner in which he ought not to act. And so it follows from the Commitment Transmission Principle that Ascot ought not to intend to take the bale on the left. A similar explanation applies in Thinking Outside the Box, where intending to take box B would commit Boxer to not reopening deliberation, when she ought to do otherwise. Thus, because intentions commit us to more than just the act intended, the Commitment Transmission Principle can explain why object-given reasons, or features of the act intended, are not the only considerations bearing on what we should intend.

Consider, finally, affective attitudes such as fear, gratitude, or anger. Suppose I'll be killed unless I fear a harmless mouse, or unless I am angry toward someone who has done nothing wrong, or unless I am grateful toward someone who has done nothing for me. The commitments principle explains why, in spite of the rewards, I ought not to have these attitudes. For such affective attitudes likewise involve practical commitments. These attitudes, like intentions, preferences, and beliefs, serve as surrogates for deliberation. In response to evidence that some object is dangerous, we come to fear the object. And we can then simply rely on this fear to guide our conduct in contexts of action when we may not have the luxury of deliberating about whether to avoid the object. Thus, if we have come to fear snakes, and if on some occasion a snake crosses our path, our fear will suffice to guide us to stop in our tracks, without the need to weigh the pros and cons of so acting. And so it follows from the criterion of commitment we proposed earlier that fearing an object commits one to avoiding it. Now if someone will kill me unless I fear a harmless mouse, I will nonetheless lack sufficient reason to avoid the mouse. Hence, I will lack sufficient reason to do what the fear commits me to doing. Thus, by the Commitment Transmission Principle, I ought not to fear the mouse. Similarly, if someone will kill me unless I am angry at a moral saint, I will nonetheless lack sufficient reason to act toward the moral saint in the manner in which my anger toward him would commit me to acting. And so, again, the Commitment Transmission

Principle implies that I ought not to be angry toward the moral saint. Thus, the Commitment Transmission Principle can explain why we ought not to have irrational affective attitudes.

The Commitment Transmission Principle can also explain a peculiar feature of affective attitudes, which competing views have difficulty explaining. It is commonly thought that a reason to fear something must be a reason to believe that it is dangerous, that a reason to be grateful toward someone must a reason to believe this person benefitted one, that a reason to be angry at someone must be a reason to believe this person wronged one, and, more generally, that a reason to have an attitude must be a reason to believe that this attitude is objectively appropriate or fitting. But this leaves unanswered the question of *how much* reason one must have to think an attitude is objectively fitting or appropriate in order for it to be the case that one ought to have this attitude, or in order for this attitude to be rational. And it seems the answer to this question can vary significantly from case to case and from attitude to attitude, and that, sometimes, it can be rational to have an attitude even when one should think that it is *probably not* objectively fitting or appropriate. Thus, if I think there is a small chance that a snake is dangerous, but that if the snake is dangerous it is *very* dangerous, then it would be rational for me to fear the snake. The same is not true of anger: if I think there is a small chance that someone wronged me, but that if he wronged me he wronged me very severely, it would not be rational for me to be angry at him. The Commitment Transmission Principle provides a nice explanation of this. In the first situation, I ought (in the deliberative sense) to avoid the snake, and so I ought to act as my fear of the snake commits me to acting, but in the second case I ought not (in the deliberative sense) to act retributively toward the person, and so I ought not to act as my anger toward him would commit me to acting. In general, attitudes can differ in the level of evidence they require to make them rational, because they can differ in the level of evidence required to make it the case that we ought (in the deliberative sense) act in the ways to which these attitudes would commit us.

6.3 Why We Ought Not to Have Irrational Combinations of Attitudes

In the last section I argued that we ought not to have irrational attitudes, on the ground that such attitudes commit us to acting in ways in which we ought not to act—that is, they commit us to acting in ways for which we lack sufficient reason, relative to our evidence. And so it follows, by the Commitment Transmission Principle, that we lack sufficient reason for these attitudes, or in other words that we ought not to have them. In the present section I will argue, similarly, that we ought not to have irrational *combinations* of attitudes. To do so, I will argue, on the basis of the Commitment Agglomeration Principle, that irrational combinations of attitudes commit us to acting in ways for which we lack sufficient reason. Hence it will follow, from the Commitment Transmission Principle, that we lack sufficient reason to have such combinations of attitudes, even if we have sufficient reason to have each of the individual attitudes of which they consist. Thus, the Commitment Agglomeration Principle, together with the Commitment Transmission Principle, will allow us to solve the problem of mere incoherence.

Now in order to have sufficient reason to ϕ in C, two things are necessary: first, ϕ-ing must be possible (i.e., it must be that if one were in C, one would be able to ϕ), and, second, it cannot be the case that in C there is some alternative to ϕ-ing for which one has more reason. Thus, there are two ways in which a mental state may commit one to doing something one lacks sufficient reason to do: it may commit one to doing something one cannot do, or it may commit one to doing something than which some alternative would be better. Any irrational combination of attitudes, I will now argue, involves one or other of these problematic commitments.

Let us first consider combinations of attitudes that commit one to doing what is impossible. This is true of inconsistent intentions. Suppose one has inconsistent intentions in a Buridan's Ass case: one intends to take only the bale of hay on the left, and one also intends only to take the bale of hay on the right. This is a decent candidate for being a case of mere incoherence: for here it seems that one has sufficient reason for each intention taken on its own, and yet together they are incoherent. But here it is clear that this combination of intentions commits one to doing something impossible, namely (taking only the bale on the left and taking only the bale on the right). For the first

intention commits one to taking only the bale on the left, and the second commits one to taking only the bale on the right. And so, by the Commitment Agglomeration Principle, the mental state consisting in the pair of intention commits one to the conjunction of these ways of acting.

As another illustration of a combination of attitudes that commits one to an impossible course of action, consider intransitive preferences. Suppose I prefer chocolate ice cream to vanilla, vanilla to strawberry, and strawberry to chocolate. This combination of preferences commits me to doing the impossible, namely to taking none of the options when taking chocolate, taking vanilla and taking strawberry are my *only* options (i.e., in circumstances in which I don't have the option of declining every flavor of ice cream). For the preference for chocolate over vanilla commits me to not taking vanilla in any circumstance where my only options are taking chocolate, vanilla, or strawberry. And the preference for vanilla over strawberry commits me to not taking strawberry in any circumstance where my only options are taking chocolate, vanilla, or strawberry. Similarly, the preference for strawberry over chocolate commits me to not taking chocolate in any circumstance where my only options are taking chocolate, vanilla, or strawberry. Hence, by the Commitment Agglomeration Principle, a mental state that consists in having all three of these attitudes will commit me to not taking any of my three options in any circumstance where my only options are taking chocolate, vanilla, or strawberry. And so such a complex mental state will commit me to doing the impossible.

It seems, therefore, that inconsistent preferences and inconsistent intentions both commit one to doing what one cannot do. And so it follows from the Commitment Transmission Principle that one lacks sufficient reason for incoherent attitudes of either kind. Other incoherent sets of attitudes commit one to acting in ways which, though possible, are inferior to some alternative. This is true, in particular, of incoherent beliefs or incoherent credences. For anyone with such incoherent attitudes will be vulnerable to a Dutch book, and will thus be committed to taking every bet in a set of bets which together would result in a sure loss. As a simple illustration, suppose I have credence .7 that when a given die is cast it will not come up 1 or 2, and I also have credence .7 that it will not come up 3 or 4, and I similarly have credence .7 that it *will* come up 1 or 2 or 3 or 4. The first of these credences commits me to taking a bet that costs $.70 and pays a

dollar if the die doesn't come up 1 or 2; the second credence commits me to taking a bet that costs $.70 and pays a dollar if the die doesn't come up 3 or 4; and the third credence commits me to taking a bet that costs $.70 and pays a dollar if the die does come up 1 or 2 or 3 or 4. Hence, by the Commitment Agglomeration Principle, the mental state consisting of all three credences commits me to taking all three of these bets. But taking all three of these bets would cost me $2.10, and is guaranteed to return only $2, and so it would result in a sure loss. Thus, the mental state consisting in all three credences commits me to acting in a certain way (namely, accepting all three bets) when I have more reason, relative to my evidence, to act in some alternative way (namely, declining all three bets). And so this combination of attitudes commits me to acting in way in which I ought, in the deliberative sense, not to act.

One criticism of Dutch book arguments is that they are said to provide the wrong kind of reasons for avoiding incoherence. All they show, it is argued, is that there are *pragmatic* reasons for avoiding incoherent credences. Hence, all they show is that we have reason to want, and to strive, not to have incoherent credences. Such arguments do not show, it is claimed, that incoherent credences are *irrational*, any more than the fact that one would be tortured for believing that 1 = 1 shows that it would be irrational for one to believe that 1 = 1.

But this argument fails to take into account an important difference between two kinds of practical consideration. In particular, it fails to distinguish between a mental state's *having negative side effects* and a mental state's *committing one to doing* something one ought not to do. Clearly, a mental state can be perfectly rational in every respect and yet have negative side effects, as the case of Floyd illustrates. It doesn't follow, however, than a mental state can commit one to doing something that one ought not to do, and yet be perfectly rational.

I have argued that, on the basis of the Commitment Agglomeration Principle together with the Commitment Transmission Principle, we can explain why we ought not to have irrational combinations of attitudes, regardless of whether we have sufficient reason for each of the constituent attitudes taken on its own. One might object, however, that this explanation fails to address the initial grounds for denying that this could be so. The worry, recall, was that it seems it can only be the case that we ought to ϕ if we have

reason to ϕ, and a reason to ϕ must be a reason *for which* we could ϕ, i.e., it must be the kind of consideration *on the basis of which* we ϕ. And yet, one might assume, the only considerations on the basis of which we could have or lack a combination of attitudes must be reasons for or against the *individual attitudes*.

But we are now in a position to argue for the denial of this last assumption. That is, we are in a position to argue that our reason for having or avoiding a combination of attitudes needn't be a reason for or against any of the individual attitudes. In particular, our reason for avoiding a combination of attitudes might be that it would commit us to doing what we ought not to do. This is true because we can engage in the practical analogue of *reductio ad absurdum* reasoning. In ordinary *reductio* reasoning, we adopt a set of suppositions, and then reason hypothetically on their basis, until we arrive at a conclusion that we clearly ought not to believe. Arriving at such a hypothetical conclusion will motivate a rational agent not to accept all the premises from which the argument proceeded, even if she cannot find fault with any one of them taken on its own. In the practical analogue of *reductio* reasoning, we suppositionally adopt a set of attitudes, and then reason practically on their basis, until we arrive at a practical conclusion that is clearly unacceptable. Arriving at such a hypothetical conclusion will motivate a rational agent not to adopt all the attitudes from which the reasoning proceeded, even if she can't find fault with any one of these attitudes taken on its own. The fact that that we can be under rational pressure not to have a given set of attitudes in virtue of the unacceptable *practical commitments* it would involve is no more mysterious than the fact that we can be under rational pressure not to have a given set of beliefs in virtue of the unacceptable *theoretical commitments* it would involve.

6.4 Further Challenges

I have argued that we ought, in a normative sense, *not* to have irrational attitudes or combinations of attitudes. But there is more to rationality than avoiding irrational attitudes and combinations of attitudes. In addition to requiring us *not* to have certain attitudes, rationality can require us to *have* certain attitudes (e.g., the belief that one exists or that 1 = 1, or the preference for a smaller pain on Wednesday to a greater pain on Tuesday). And in addition to requiring us *not* to have certain combinations of attitudes,

rationality can require us to be such that, *if* we have certain attitudes, *then* we also have others. For example, rationality can require us to be such that, if we believe that Socrates is a man and that all men are mortal, then we believe that Socrates is mortal. And rationality can require us to be such that, if we intend to bake a cake and believe that buying sugar is a necessary means to baking a cake, then we intend to buy sugar. Finally, rationality can require not only that we have certain attitudes, but also that we have these attitudes for certain reasons, and hence that the appropriate grounding relations exist among our attitudes. A complete vindication of the normativity of rationality would therefore have to show that we ought, in a normative sense, to satisfy all these kinds of rational requirement. Such a complete vindication of the normativity of rationality, however, exceeds the scope of this paper.

References

Arntzenius, Frank, Adam Elga and John Hawthorne (2004) "Bayesianism, Infinite Decisions, and Binding" *Mind* 113: 251-283.

Broome, John (unpublished) *Reasoning*. Unpublished lectures delivered at Brown University and the University of Stockholm.

Brunero, John (unpublished) "'Ought' and the Perspective of the Agent," unpublished manuscript.

Finlay, Stephen (2009) "Oughts and Ends" *Philosophical Studies* 143: 315-340.

Finlay, Stephen (2009) "What *Ought* Probably Means, and Why You Can't Detach It" *Synthese* 177:67-89.

Goldman, Holly (1976) "Dated Rightness and Moral Imperfection," *Philosophical Review* 85 (4): 449-487.

Hieronymi, Pamela (2005) "The Wrong Kind of Reason" *Journal of Philosophy* 102: 437-457.

Kavka, Gregory (1983) "The Toxin Puzzle" *Analysis* 43: 33-36.

Kolodny, Niko (2005) "Why Be Rational?" *Mind* 114: 509-563.

Kolodny, Niko (2007) "How Does Coherence Matter" *Proceedings of the Aristotelian Society* 107: 229-263

Oddie, Graham and Peter Menzies (1992) "An Objectivist's Guide to Subjective Value" Ethics 102: 512-533.

Parfit, Derek (1984) *Reasons and Persons*, Oxford: Oxford University Press.

Parfit, Derek (1997) "Reasons and Motivation" *Aristotelian Society Supplementary Volume* 77: 99-130.

Parfit, Derek (2001) "Rationality and Reasons" in Dan Egonsson *et al*, eds. *Exploring Practical Philosophy*, Burlington: Ashgate, 17-39.

Parfit, Derek (2011) *On What Matters*, Oxford: Oxford University Press.

Piller, Christian (2001) "Normative Practical Reasoning" *Proceedings of the Aristotelian Society*, Suppl. Vol 25: 195-216.

Raiffa, Howard (1970) *Decision Analysis*, Reading, Mass: Addison-Wesley.

Raz, Joseph (2005) "The Myth of Instrumental Rationality" Journal of Ethics and Social Philosophy: 2-28.

Regan, Donald (1980) *Utilitarianism and Cooperation*, Oxford: Oxford University Press.

Ross, Jacob (2006) *Acceptance and Practical Reason*, doctoral dissertation, Rutgers University.

Ross, Jacob (2010a) "The Irreducibility of Personal Obligation" *Journal of Philosophical Logic* 39: 307-323.

Ross, Jacob (2010b) "Sleeping Beauty, Countable Additivity, and Rational Dilemmas" *Philosophical Review* 119: 411-447.

Ross, Jacob (forthcoming) "Actualism, Possibilism, and Beyond" possibly forthcoming in *Oxford Studies in Normative Ethics*, volume 2.

Ross, Jacob and Mark Schroeder (forthcoming), "Belief, Credence, and Pragmatic Encroachment." Forthcoming in *Philosophy and Phenomenological Research*.

Scanlon, T. M. (1999) *What We Owe to Each Other*, Cambridge: Harvard University Press.

Scanlon, T. M, (2007) “Structural Irrationality” in *Common Minds: Themes from the Philosophy of Philip Pettit*, Brennan, Goodin, Jackson and Smith, eds., New York: Oxford University Press.

Schroeder, Mark (2008) “Having Reasons” Philosophical Studies 139: 57-71.

Schroeder, Mark (forthcoming) “What Does it Take to ‘Have’ a Reason,” forthcoming in Reisner and Steglich-Peterson (eds.) *Reasons for Belief*.

Way, Jonathan (2009) “Two Accounts of the Normativity of Rationality” *Journal of Ethics and Social Philosophy*, December 2009.

Wedgwood, Ralph (2003) “Choosing Rationally and Choosing Correctly” in *Weakness of Will and Practical Irrationality*, edited by Sarah Stroud and Christine Tappolet. New York: Oxford University Press.

Williamson, Timothy (2002) *Knowledge and Its Limits*. Oxford: Oxford University Press.