

## Instrumental Rationality

Ralph Wedgwood

**0.** Is there any distinctive aspect of rationality that deserves the label of “*instrumental rationality*”? Recently, Joseph Raz (2005) has argued that instrumental rationality is a “myth”. In this essay, I shall give some qualified support to Raz’s position: as I shall argue, many philosophers have indeed been seduced by certain myths about instrumental rationality. Nonetheless, Raz’s conclusion is too strong. Instrumental rationality is not itself a myth: there really is a distinctive aspect of rationality that deserves the label of “instrumental rationality”.

In the first two sections of this essay, I shall start by giving a rough intuitive description of the phenomenon that seems to me the best candidate for the label “instrumental rationality”. As we shall see, this rough description gives us reason to reject some of the myths that surround instrumental rationality. Then in the rest of this essay, I shall try to give a more precise general specification of this phenomenon. In Sections 3 and 4, I shall consider what has been said about instrumental rationality by several other philosophers. Identifying what is missing in these other philosophers’ accounts will help me to develop my own positive specification, which I shall present in Sections 5 and 6.<sup>1</sup>

**1.** Let us start with an example of the kind of reasoning that it is most natural to call “instrumental reasoning”. Suppose that I have decided to pursue an *end* or *goal* (I shall not distinguish between these terms here). For example, suppose that my goal is to walk to the top of Snowdon today.

---

<sup>1</sup> So my aim in this essay is simply to give a precise general specification of this aspect of rationality, and to argue that this specification is correct. It is *not* my aim here to look for the most fundamental explanation of this aspect of rationality – let alone to investigate the nature of rationality itself, or to inquire into the ultimate sources of rational norms or the like. These are excellent questions – indeed, crucial questions – but I shall not be pursuing them here. It may be helpful to start with a precise *identification* of a phenomenon before setting out to find the *explanation* for it.

Instrumental reasoning will take me from having this goal to thinking about *how* more exactly to achieve this goal.

For example, I might think about which of the various different *routes* to take to the top of Snowdon. There is the path from Llanberis in the North. But that would not be much of a challenge, and I would be surrounded by crowds of tourists all the way. There is the Crib Goch route from the East; but that path involves scrambling over a narrow rocky ridge, with precipitous cliffs on either side, and there is enough mist on the high ground to make that route seem inadvisable. So I decide on the Watkin Path from Nantgwynant in the South, which will be more isolated and challenging than the Llanberis path from the North, but less dangerous than the Crib Goch route from the East. I might go through this sort of process several times: first, I decide which route to take; then I might decide what time to do the walk, and then what equipment to take with me. At every stage, I am willing to go back and revise the decisions that I have reached so far, in view of any new decisions that I have made, and new information or considerations that have come to light.

At first glance, this process of reasoning seems to involve the following steps:

- (a) I have an *intention* to achieve an *end*.
- (b) I form beliefs about what means are *available* for achieving that end.
- (c) I form beliefs about which of these means are *better*, and which are *worse*, than others.
- (d) Finally I *choose* one of these available means that I believe to be *optimal* (i.e. no worse than any other) – or at least, means that I do *not* believe to be *worse* than any other.

In what follows, I shall comment on each of these four steps in turn.

**a.** As I have described it, this kind of reasoning starts out from an *intention* to achieve an end. An intention is different from a mere desire: resolute ex-smokers may be tormented by the desire to smoke, but thanks to their will-power, never form an intention to smoke. Of course, there could also be a different sort of reasoning – which starts, not from an intention to achieve the end, but from a mere desire for the end, or from the belief that it would be in some way desirable to achieve the end,

or the like. Indeed, yet another sort of reasoning just starts from the conception of a possible end (which one may not desire or regard as worth pursuing at all), and proceeds to inquire into the best way of achieving that end (for example, one might engage in this sort of reasoning purely in a spirit of idle curiosity, out of a purely theoretical interest in discovering what would be the best way of achieving the end in question).

In this essay, however, I shall focus entirely on the kind of instrumental reasoning that starts from an intention to achieve the end. An intention to achieve an end, a desire for the end, and a mere conception of the end are mental states of three radically different kinds; so it is surely plausible that they are subject to equally different requirements of rationality. An intention involves *committing oneself* to achieving the intended end, while neither a desire for the end nor the mere conception of the end involves committing oneself in this way; so it seems likely that intentions will be subject to more demanding requirements of rationality than desires. Although I shall touch briefly on the rational significance of desires later, our concern here is to see if there is any distinctive aspect of rationality that deserves the label “instrumental rationality” – that is, an aspect of rationality that is especially connected with the kind of reasoning that takes us from ends to means. Since intentions are subject to more demanding requirements of rationality than desires, it seems likely that this aspect of rationality will be particularly evident in the kind of reasoning that starts from intended ends (rather than ends that are not intended), and results in our forming an intention to take some appropriate means.

**b.** In my description, the second step in this process of instrumental reasoning, after acquiring an intention to achieve the end, is to form beliefs about which means to that end are *available*. But what is it to believe that a certain course of action is “available” in the relevant way?

I propose that believing that a certain course of action is available in this way has two components. First, one must regard it as at least epistemically possible that one will intend that course of action – that is, in effect, one must attach a non-zero probability to one’s intending that course of action. Secondly, one has a confident conditional belief that one will in fact take that course of action if one intends to do so; that is, in effect, one attaches a high *conditional probability* – in the context, amounting for all practical purposes to conditional certainty – to the hypothesis that one will take that

course of action, given the assumption that one intends to do so. As we might put it, the sort of availability that one believes these courses of action to have is, more specifically, *availability through intention*.

This is admittedly a demanding conception of what it is to believe an option to be available. On this conception, it is unlikely to be rational for a habitual smoker to believe that the option of *giving up smoking* is available (since the smoker is nowhere near conditionally certain that he will give up smoking if he intends to do so); at most, it would be true to say that it is rational for the smoker to believe that *trying to give up smoking* is available. So the view that this demanding conception is the appropriate one to use in an account of instrumental rationality requires some further justification and defence; I shall provide some further defence for this view in Section 3b.

Although instrumental reasoning involves identifying *some* means that are available for achieving the end, it need not involve identifying *every* available means to the end. Indeed, in some cases, the reasoner may just identify *one* available means to the end. This will be particularly common when the reasoner takes it for granted that none of the available means will be any better or any worse than any other. In such cases, there will be no need to compare alternative means with each other; so it will be enough to identify just one means to the end.

c. The third step in my description of instrumental reasoning is to form beliefs about which of these possible means for achieving the end are *better* than others. In instrumental reasoning, one is focusing solely on evaluating various possible means for achieving one's end – that is, various courses of action that could in some way or other facilitate the achievement of this end. Still, even if one restricts one's attention to means to this end, one still needs to judge how good or valuable these means are overall. In effect, one needs to judge how "*choiceworthy*" each of these available means is – that is, how worthy of being chosen each of these available means is.<sup>2</sup>

Some philosophers may reject this conception of the third step. According to these

---

2 I have discussed this notion of "choiceworthiness" in several places elsewhere; see e.g. Wedgwood (2003, 211; 2004, 421; 2007, 101-04).

philosophers, in instrumental reasoning, one is purely concerned with evaluating the various possible means in terms of their *effectiveness* at facilitating the goal. Even if one expresses one's evaluation of a pair of possible means by saying one of these means is a "better" means of accomplishing the goal than the other, all that one means by this is that the first course of action is *more effective* at accomplishing the goal, or that the first makes the goal *more likely* to be accomplished than the second, or the like.

On closer inspection, however, the idea that we often choose a means to an end purely on the basis of its effectiveness at facilitating the end seems to be a myth. In fact, we seem rarely to make any but the most trivial choices in this way. For example, when I was reasoning about how to climb Snowdon, I decided that the Watkin path from the South was better than the Llanberis path from the North. But this was not because the Llanberis path is in any way less effective as a way of walking to the top of Snowdon than the Watkin path. On the contrary, it is, if anything, *more* effective. (It is, after all, a considerably easier and less challenging route.) The Watkin path is a better way for me to walk to the top, *not* because it is a more effective way for me to walk to the top, but because it does better at satisfying *all* of the many values and desiderata that can be satisfied by any of the various possible ways of walking up Snowdon. Indeed, quite generally, the most effective way of accomplishing a goal will frequently be too costly or disagreeable or boring or painful or in some other way objectionable, to count as the best, or as even one of the best, ways of accomplishing the goal.

Moreover, it would in a great many cases be utterly irrational to form an intention to take whatever one believed to be the most effective means for achieving one's goal. If it is obvious that the most effective means would be ruinously expensive, or unbearably tedious, or morally hideous, then it would be quite irrational to ignore these powerful reasons against taking the most effective means. If one is to be rational, one needs to take all of these reasons into account. While it might be an interesting theoretical exercise to work out what the most effective means to a certain end would be, if instrumental reasoning is a form of *practical* reasoning, it can be rational only if it is sensitive to *all* the relevant reasons, values, and desiderata that bear on whether or not the available means are worthy of being chosen.

Some philosophers may concede that the best way of achieving a goal *E* may not always be

the means that are most effective for achieving *E*, but they might insist that the “best way of achieving *E*” can still be defined as a way of achieving *E* that is the most effective means of promoting *all* of the relevant ends – where the “relevant ends” include several other ends besides *E*. But precisely which ends are these “relevant ends”?

One possible interpretation is that these “relevant ends” are simply the ends that the agent *already intends* to achieve. In effect, according to this interpretation, the only factor that can make one means to an end *E* “better” than another means is the extent to which these means promote the ends that the agent already intends to achieve. But again, this seems not to be how we normally think. Whenever one intends to achieve an end, one has *committed* oneself to achieving that end. When I judge that the Watkin path is a better way to the top of Snowdon than the Llanberis path, it need not be that every feature of the Watkin path that leads me to judge it to be better consists in its promoting an end that I have already definitely committed myself to achieving, in advance of deciding to take that path.<sup>3</sup> So it seems that this notion of the “best way of achieving *E*” is not simply the notion of the means that are most effective at promoting the ends that the agent already intends to achieve.

An alternative interpretation would take the “relevant ends” to be the objects of the relevant agent’s *desires*. In effect, this interpretation implies that one way of achieving an end *E* counts as “better” than another way of achieving *E* if and only the first way of achieving *E* satisfies *the agent’s desires* more effectively than the second. However, this interpretation may well be *entirely compatible* with the claims that I have made.

For the purposes of this discussion, I have not taken any stand on what exactly the nature of “*choiceworthiness*” is. All that matters for my purposes is that there is a kind of judgment that compares and evaluates the courses of action that are available to a given agent at a given time, which

---

<sup>3</sup> Moreover, even if the agent does already intend to achieve all of these relevant ends, how exactly should these different ends be *balanced* against each other? Again, I think that the only plausible answer is to say that these ends should be balanced against each other in the *right* or *appropriate* way – the way in which they would be balanced by someone who was deliberating correctly about how to achieve the relevant end. So appealing to a richer evaluative or normative notion (such as “choiceworthiness” or the like) seems indispensable here too.

can be expressed by calling some of these courses of action “better” or “more choiceworthy” than others, and which reflects an attempt to be sensitive to the overall upshot of *all* the various values and reasons for action that are at stake in the agent’s situation at that time.

It may well be that these judgments are indeed closely connected to desires. One suggestion is that the connection is with the *truth conditions* of these judgments: the judgment that one course of action is “better” than another is true if and only if the first course of action satisfies the agent’s overall set of desires more effectively than the second. Another suggestion is that the connection is with what makes these judgments *rational*: it is rational for a thinker to judge that one course of action is better than another if and only if this judgment is supported in the appropriate way by the thinker’s desires.<sup>4</sup> However, I do not need to assess either of these suggestions here. The only assumption that we need here is that there is a kind of judgment, expressible by terms like ‘better’ and ‘more choiceworthy’, that is sensitive to all the reasons that the relevant agent has for and against the courses of action that are available at the relevant time. We can remain agnostic here about how exactly (if at all) choiceworthiness is connected to desires.

In the description that I have given, the second step of instrumental reasoning consists of identifying certain means to the end as available, and the third step consists in comparing and evaluating these means, by forming beliefs about which of these means are better than others. So the third step is only concerned with comparing the means that have actually been identified as available at the second step. In the simple case where one has identified only *one* means, it is in effect *vacuously* true that this means is not worse than the others: in this case, the phrase ‘the others’ picks out the empty set, and nothing can be worse than any of the members of the empty set. In these cases, then, it will be particularly easy to form a belief about how the means that one has identified as available compares with the alternative means (if any) that one has identified.

---

4 As a matter of fact, I would argue against the first of these two suggestions (see Wedgwood 2002) and in favour of the second (see Wedgwood 2007, Chapter 10). But for the purposes of this essay, I do not need to make this argument here.

d. The final step in this description of instrumental reasoning is relatively straightforward: it is to *choose* – that is, to come to intend – one of the possible means that one has identified as available – and specifically to choose one of the means that one believes to be optimal, or at all events to choose one of these means that is *not* such that there is an alternative that one believes to be *better* or *preferable* to it.

So far I have spoken repeatedly of *means* to an *end*. But I think that the fundamental notion here is that of *taking* a course of action *as* a means to an end. (The various possible means that one evaluates when engaged in instrumental reasoning are all courses of action each of which one might possibly take *as* a means to the end.) To take a certain course of action *A* as a means to a certain end *E* is to carry out an *intention* to take that course of action *A in order to* achieve that end *E*. The “in-order-to” relation, as I am understanding it here, need have nothing to do with the ultimate *motivation* or *justification* of one’s intentions: it has to do with the *structure* of one’s intentions. When one intends to take a course of action *A in order to* achieve an end *E*, one’s intention to take *A* is in a way subordinate to one’s intention to achieve *E*. Roughly, one’s intention to achieve the end *E* “controls” or “guides” the way in which one takes the course of action *A*: in executing one’s intention to take this course of action *A*, one will continually monitor one’s conduct and adjust it in such a way as to facilitate the achievement of *E*.<sup>5</sup>

Presumably, one possible means counts as an “alternative” to another possible means only if it is impossible to take both. (For example, the Llanberis path and the Watkin path are alternative means of walking to the top of Snowdon, since it is impossible to walk exactly once to the top of Snowdon along both paths.)

According to the description that I have given so far, the third step consists in forming *explicit* beliefs about which of the relevant means are better than others, and the fourth step consists of choosing a means that one believes not to be worse than any alternative (or at least a means such that there is no alternative that one believes to be better than it). But instrumental reasoning is often much less explicit than this. So it seems that this description should be relaxed, so that it no longer requires

---

5 For this idea of an intention’s “controlling” one’s behaviour, see especially Mele (2000).

forming explicit beliefs in this way. It is enough if the reasoner's *choices* are *sensitive* to the evidence or reasons that support or justify such beliefs about which of the available means are better and which are worse. Sensitivity of this kind would require that the reasoner chooses an available means *A* only if the reasons and evidence at the reasoner's disposal justify the belief that *A* is no worse than any of the alternatives (or at least do not justify any belief, concerning an alternative *B*, to the effect that *A* is worse than *B*). It seems to be possible for a reasoner's choices to be sensitive to reasons and evidence in this way even if the reasoner does not form any explicit beliefs about which courses of action are better than any alternatives.

This then is a rough description of what instrumental reasoning involves. Instrumental *rationality* consists, presumably, in doing such instrumental reasoning *rationally*. In the next section, I shall try to give a description of at least some of the conditions that must be met if one is to do this sort of instrumental reasoning in a rational manner.

2. Roughly, it seems that someone would be doing instrumental reasoning in a rational manner only if (a) they intend to achieve an end, and respond to the evidence that they have by forming (b) *rational* beliefs about what means of achieving that end are available, and (c) *rational* beliefs about which of these means are better and which are worse, and (d) intentions that are (as we might put it) *in line with* these rational beliefs.

In general, there are at least two ways of forming intentions that are "in line with" one's beliefs about which are the best means to this end: roughly, either (i) one can come to intend some course of action that one rationally believes to be one of the available and optimal means to the end (or at least some means such that there is no alternative to which one believes it to be inferior), or (ii) one can drop the intention to achieve the end. In many cases, there will be compelling reasons of some kind in favour of one of these two ways of keeping one's intentions in line with one's beliefs (and against the other way of keeping them in line with each other). But from the point of view of instrumental rationality, the two seem on a par: neither is more or less instrumentally rational than the other.

To put it roughly again, instrumental irrationality must consist in some corresponding *failure* of instrumental rationality. So it seems that one way in which one may be instrumentally irrational is if one's intentions are not in line with one's beliefs in this way. So, roughly, it is irrational simultaneously to intend an end, and to believe, of a certain set of alternative means, that each of them is an available and optimal means to the end, while intending none of these means. Strictly, this is not quite precise, since it surely need not be irrational to *postpone* making up one's mind about some practical questions until it becomes necessary to do so. So perhaps we should say, to more precise, that it is irrational to *persist* with not intending any of these means if one also rationally believes that one will not achieve the end in an optimal way unless one now decides to take one of these means.

This characterization of this sort of instrumental irrationality is designed to cover, not only those cases in which there is a *plurality* of means none of which one believes to be worse than any alternative, but also the cases in which there is a course of action that one believes to be the unique *best* means to the end – since, presumably, if a rational person believes a course of action to be the unique best means to an end, that means will be the only one that the person believes to be an optimal means to the end. Similarly, it seems that if a rational person believes a certain course of action to be the *only* possible means to the end, that course of action will again be the only one that the person believes to be an optimal means to the end. Thus, this formulation also covers the case in which one believes that a certain course of action is the only possible means to the end.

In effect, this characterization of instrumental irrationality articulates what has come to be called a “wide-scope” rational requirement. That is, the distinctively instrumental irrationality is not primarily located in any one mental state, but rather in the *combination* of intending the end, believing that a certain set of means are the optimal means to that end, and not intending any of those means. There need be no irrationality at all – and certainly no distinctively *instrumental* irrationality – in any one of these mental states all by itself: the instrumental irrationality crucially consists in this combination of mental states.<sup>6</sup>

---

<sup>6</sup> The wide-scope reading of such requirements of instrumental rationality has been defended by Broome (1999), and criticized by Kolodny (2005), Raz (2005), and Schroeder (2004). In my view, on this point Broome

It is important that this requirement of instrumental rationality is restricted to *means*. One is not in general rationally required to intend everything that is necessary for the achievement of one's end – let alone every component of what will bring about the optimal achievement of the end. For example, I believe that in order for me to reach the top of Snowdon today it is necessary that I do not undergo spontaneous human combustion while walking up Cwm Llan. However, even though I have this belief, and also intend to get to the top of Snowdon, there need be nothing irrational in simultaneously lacking any intention not to undergo spontaneous combustion. This is because I also believe that I will not undergo spontaneous combustion, quite irrespective of whether or not I intend not to. This belief seems to make it impossible for me to intend not to undergo spontaneous combustion. So, avoiding spontaneous combustion is not in my sense a *means* to getting to the top of Snowdon: it is not something that I can intend in order to get to the top.

Even though the requirement of instrumental rationality that I have just sketched is restricted to means in this way, it is tempting to think that there are other rational requirements that should also be captured by a good theory of instrumental rationality. In particular, it is tempting to think that in addition to the requirement that one should intend appropriate means to one's ends, there is also a requirement that we should *not* intend states of affairs that one believes will prevent the achievement of the ends that one intends. This is in effect the requirement that Michael Bratman (1987, 31) calls the “demand that my plans should be strongly consistent, relative to my beliefs.” Bratman formulates this demand as follows: “Roughly, it should be possible for my entire plan to be successfully executed, given that my beliefs are true.” In Section 5, I shall try to find to an account of instrumental rationality that can capture Bratman's requirement of “strong consistency” as well as the requirement that one should intend some optimal means to one's end.

In the previous section, I conceded that there could be cases where the reasoner only identifies *one* course of action as an available means to the end. As I explained, in these cases, it is *trivial* to form a rational belief to the effect that this course of action is not worse than any of “the alternatives”

---

is right, and the criticisms of Kolodny, Raz, and Schroeder are entirely mistaken. But I shall not have time to answer these criticisms here.

(since “the alternatives” in this context is simply the empty set). So the requirement of instrumental rationality that I have described here is particularly easy to meet in the case where the reasoner only identifies one course of action as an available means to the end.

However, this gives rise to a question: Might it not sometimes be irrational to consider *too few* alternatives? (And might it not also sometimes be irrational to waste time by considering *too many* alternatives?) The answer to this question is surely: Yes, there are indeed cases in which it is irrational to consider too few alternatives in this way (and there are also cases in which it is irrational to consider too many alternatives). Unfortunately, however, I shall not be able to explain exactly why it is irrational to consider too few alternatives in these cases. For this reason, the account that I have given here should only be taken as a statement of a *necessary* condition of rationality – not as a statement of a *sufficient* condition of rationality. At all events, the requirement that we should not consider too few alternatives in these cases is not a special feature of *instrumental* rationality: it would equally apply to the sort of practical reasoning that involves deciding on ultimate ends – since this sort of reasoning could also in many cases count as irrational if one decides on an ultimate end without considering sufficiently many alternatives.

Even if it is taken merely as a statement of a necessary condition of rationality, the account that I have given so far is rough and imprecise in several ways. There are two points where the rough and imprecise character of the description that I have given so far seems particularly troubling:

- (a) Some courses of action are *parts* or *components* of larger courses of action. How does this description accommodate this point? Could not some rational intentions be for the *parts* or *components* of some larger course of action – while it is only that larger course of action (not its smaller parts or components) that one rationally believes to be one of the optimal means to the end?
- (b) What about occasions when we are *uncertain* about crucial features of our situation? On such occasions, are we really in a position to form any rational beliefs about which of the available options are better and which are worse? If so, how exactly should we think of such “beliefs”?

For the rest of this essay, I shall try to give a more accurate and more complete account of instrumental rationality that will clarify these two points. In the next two sections, I shall start by considering what some other philosophers have said about the topic.

**3a.** The first view of instrumental rationality that I shall canvas here is the view of Joseph Raz (2005). As I have already mentioned, Raz thinks that instrumental rationality is in a sense a “myth”. According to Raz, the only real phenomenon in this area is what he calls the “facilitative principle” – namely, the principle that if one has sufficient reason to pursue an end, one also has a reason to take any course of action that facilitates that end. Apart from the facilitative principle, however, there is no other real phenomenon lying behind what philosophers have called “instrumental rationality”.

In fact, as John Broome (2005) has pointed out, Raz’s facilitative principle is in one respect too strong. To take Broome’s example, suppose that it is lunch time and you have a sufficient reason not to be hungry during the afternoon. One course of action that will facilitate your not being hungry in the afternoon is to eat the tasty and nutritious lunch that is already on the table right in front of you; another course of action that will facilitate your not being hungry this afternoon is to kill yourself. It seems to me most doubtful whether your reason not to be hungry this afternoon generates *any reason at all* for you to kill yourself.<sup>7</sup>

However, this aspect of Raz’s principle can easily be fixed. We need only amend the principle so that it says that whenever one has a sufficient reason to pursue an end, and a course of action is an *optimal* means to that end, then one also has a reason to take this course of action. Indeed, we could strengthen this so that it says that whenever one has a sufficient reason to pursue an end, and a course of action is an optimal means to that end, then one also has a *sufficient* reason to take this course of action. (The crucial distinction here is between merely having *a reason* to take a course of action, which is compatible with that reason’s being defeated or outweighed by some reason in favour of an

---

<sup>7</sup> A similar objection applies to Dreier’s (1997, 93) “means/ end rule”: “If you desire to  $\psi$ , and believe that by  $\phi$ -ing you will  $\psi$ , then you ought to  $\phi$ .” I desire not to be hungry this afternoon, and believe that by killing myself I will not be hungry this afternoon. But it is clearly not true that I ought to kill myself!

alternative course of action, and having *a sufficient reason* – in which case the reason in question is not defeated by any countervailing reason, and the course of action in question counts as all-things-considered permissible.)

Even with this amendment, however, Raz's principle fails to account for our intuitions about instrumental rationality. Our intuition is not just that it is rationally *permissible* to intend the means to our end, but that it is rationally *impermissible* to intend an end, while believing that certain means are the optimal means to that end, without ever intending any of those means. But even if one believes (or even knows) that one has sufficient reason to take a course of action, there need be absolutely nothing rationally impermissible about not intending that course of action.

For example, suppose that I am in a "Buridan's Ass" situation with respect to both the end and the means: I only have a sufficient reason (not a compelling or overriding reason) to climb Snowdon today, since I have just as strong a reason to take some alternative course of action instead, such as spending the day climbing the Rhinog mountains to the South; and I also have only a sufficient reason to take the Watkin Path up to the top of Snowdon, since I could just as well have taken an alternative path, such as the path that goes further to the West underneath Yr Aran. So there are no courses of action that I have compelling or overriding reason to take in this case: for each of the available courses of action, I have only sufficient reason for that course of action. So if I do not take either the Watkin path or the path that goes underneath Yr Aran, the worst that Raz can accuse me of is of not doing something that I had sufficient reason to do: he cannot accuse me of not doing something that I had a compelling or overriding reason to do. But as I have claimed, there seems to be something irrational – that is, something rationally *impermissible* or *forbidden* – about the combination of intending the end of climbing to the top of Snowdon, believing that a certain set of courses of action are the optimal means to that end, and yet never intending any of those means. There seems to be no way in which Raz's principle can capture the fact that this bad combination of attitudes is rationally forbidden or impermissible in this way.<sup>8</sup> Thus, there seems to be more to instrumental rationality than Raz's account can allow.

---

8 For this sort of criticism of Raz's approach, compare Schroeder (2009, 232f.).

The same point, by the way, also shows that this requirement of instrumental rationality cannot be explained by the neo-Humean idea – which has been advocated by Bernard Williams (1981) among others – that our desires generate reasons for action. Even if my desire to achieve an end generates a reason for me to pursue that end, and so also generates a reason for me to take means to that end, this reason will in many cases only be a *sufficient* reason, not a compelling or overriding reason. In this case, even if my desires create a reason for me to take suitable means to the end, it will only be a sufficient reason, and so it will not explain why it is rationally forbidden or impermissible for me to intend the end without intending any suitable means. So the only way in which a neo-Humean approach could solve the problem would be by following Mark Schroeder's (2009) suggestion that an intention to achieve an end generates, not just *a reason*, but a *compelling* or *overriding* reason to pursue the end. But this suggestion seems completely incredible to me. Suppose that in a Buridan's Ass situation, I arbitrarily form an intention to achieve end  $E_1$  rather than the equally appealing alternative end  $E_2$ . If my intention to achieve  $E_1$  did generate a compelling reason of this kind, then it would surely also generate a compelling reason for me *not to reconsider* or *abandon* my intention to achieve  $E_1$ . But this just seems wrong: there need be nothing rationally forbidden or impermissible about my changing my mind at the last minute and deciding to achieve  $E_2$  instead of  $E_1$ . So it seems that the neo-Humean approach lacks the resources to explain this requirement of instrumental rationality.

**3b.** Let us turn now to the approaches of John Broome (1999) and Kieran Setiya (2007a). Here a quite different issue arises. In spite of several differences between their two approaches, they both share a very striking limitation: they only deal with the very special case of *necessary* means – that is, means that are strictly necessary for the achievement of the relevant end.<sup>9</sup>

In my view, no approach to instrumental rationality that applies only to the special case of strictly necessary means can count as a promising approach to the topic. First, only a tiny part of our

---

<sup>9</sup> It is possible that Kant (1785, 417) is to blame for this limitation, since that discussion of the “hypothetical imperative” also seems to be restricted to the “indispensably necessary means” to one's end.

instrumental reasoning is concerned with reasoning from an intention to achieve an end to an intention to take the *necessary* means to the end. A great deal of decision-making involves deciding between a plurality of equally possible means. Suppose that I intend to get from London to Oxford tomorrow. Then I have to decide whether to take the train or the bus or to drive. But none of these three means is necessary: I don't have to drive or take the train, because I could take the bus; but equally I don't have to take the bus, because I could drive or take the train. In all of these extremely common sorts of cases, we are reasoning to our way from an intention to achieve an end to an intention to take one of the optimal means to our end, not to take any of the necessary means to our end.<sup>10</sup>

Secondly, as I have already explained in the previous section, the requirement to intend the necessary means to one's end can be derived from the requirement to intend some optimal means to one's end. This is because any course of action that counts as a necessary means to one's end must be at least part of some course of action that is the *only* means to the end; and the only means to the end will trivially count as the optimal means to the end. So it seems that any promising approach to the issue will focus on the requirement that one should intend one of the optimal means to one's end, and will not limit itself to the very special case of means that are strictly necessary for the end.

Nonetheless, there is an important insight in Broome's and Setiya's discussions of these questions. I shall focus here on the form in which the insight is articulated by Setiya. As Setiya (2007a, 666f.) in effect points out, intending the necessary means to one's end is a necessary condition of its being rational to *believe* that one will *succeed* at achieving one's end. It seems plausible to claim that it is a necessary condition of its being rational to intend an end that it should also be rational to believe that if one has that intention, one will successfully execute the intention.

Indeed, this claim follows from what I said in the previous section about how rational instrumental reasoning involves a rational judgment about the *availability* of the various possible

---

<sup>10</sup> The means that we choose often also include *precautionary* measures, for which there is only a tiny chance of their being necessary for the achievement of the end. For example, suppose that I intend to drive to Wales without being injured in a car accident. For the sake of this end, I fasten my seat belt when I get into the car, even though I know that it is not just possible, but overwhelmingly likely, that I would still arrive safely in Wales even if I did not fasten my seat belt.

means that one could take towards one's end. According to the proposal that I made in the last section, believing that a course of action is available has two components, one of which is, in effect, the conditional belief that if one intends the option, one's will in fact execute one's intention and act accordingly. Presumably, if one does in fact intend the option in question, then one will either believe, or at least be in a position to believe, that one has this intention; and if one also has this conditional belief – namely, the belief that if one has this intention, one will successfully execute the intention – then one is committed by *modus ponens* to believing the consequent of this conditional – namely, the proposition that one will successfully execute this intention. So it does indeed seem plausible that it is a necessary condition of rationally intending an end that it should also be rational for one to believe that one will successfully execute that intention.<sup>11</sup>

However, this also reveals a further, deeper problem with both the approach of Raz and the approaches of Broome and Setiya. Raz focuses on the aspect of instrumental rationality that is concerned with beliefs about how *good* or *valuable* the available courses of action are; and as we have seen, Broome's and Setiya's proposals are most naturally explained by the aspect of instrumental rationality that concerns beliefs in the *availability* of the intended course of action. It is certainly plausible that both of these aspects are fundamental to all rational practical reasoning.

Indeed, it is not too hard to explain, at least in outline, why it is that rational practical reasoning has these two aspects. Both of these two kinds of beliefs – both a belief in the availability of the intended course of action, and a belief in this course of action's value or goodness – have the feature that unless they are *true*, the intention in question will not succeed at realizing what it is the ultimate point or purpose of intentions to realize. If the intended course of action is not available (in the sense that if one has this intention, one will execute the intention and act accordingly), then the intention will not be executed; and it is surely plausible that part of the very point or purpose of

---

11 So my position on the famous debate about whether intention implies belief (see e.g. Setiya 2007b, 46) is the following: it is not necessarily true that whenever one has an intention, one must believe that one will execute the intention and act accordingly; but it is a necessary condition on the *rationality* of an intention that it must simultaneously be rational to believe that one will execute the intention and act accordingly.

intentions that they should be executed. If the intended course of action is not good or valuable in the appropriate way, then again it seems that one has not “got things right” in one’s intentions, and so again one’s intentions will have failed to realize part of their essential point or purpose. So it is plausible that every rational intention requires some kind of rational belief in (i) the availability, and (ii) the value or goodness, of the intended course of action.

However, neither of these aspects is especially distinctive to *instrumental* reasoning as opposed to any other kind of practical reasoning. What is distinctive of instrumental reasoning is that it takes a certain end as given, and restricts its attention to the various possible means to that end. But even when one decides on an ultimate end, one must consider whether the pursuit of that end is available, and how valuable it really is. So the phenomena that Raz, Broome and Setiya focus on in their contributions are not really distinctive of instrumental reasoning.<sup>12</sup> To find what is essential to instrumental reasoning, we shall have to look elsewhere.

**4a.** One place to look for what is distinctive of instrumental rationality is to inquire whether it plays any role in formal decision theory, and if not, why not. I shall start by focusing on orthodox *causal decision theory* (CDT), of the sort that has been defended by Allan Gibbard and William Harper (1978) and by David Lewis (1981). As I shall argue, there are some fundamental idealizing assumptions built into CDT, which guarantee that CDT in fact has absolutely nothing to say about instrumental rationality. Considering how a theory of practical rationality could do without these idealizing assumptions will help us to identify the distinctive features of instrumental rationality.

Like all versions of decision theory, CDT works with some notion of value and some notion of probability, and maintains that in a sense, rational decisions *maximize expected value* – with the idea of an “expectation” being defined by means of the relevant probabilities in a broadly standard

---

<sup>12</sup> This point is not embarrassing to Raz, who is sceptical about whether there really is any such thing as instrumental rationality. But it is at least somewhat embarrassing to Setiya, who is committed to their being a distinctive realm of instrumental rationality – which for his cognitivist theory is ultimately a branch of *theoretical* rationality rather than *practical* rationality – which he is seeking to account for.

way. Many versions of CDT also embrace a broadly *subjectivist* conception of the relevant sort of value, defining it as *utility*, which is a measure of the agent's preferences (at least so long as the agent has a complete set of preferences over a large enough domain of prospects, and these preferences also meet some well-defined conditions of coherence). In fact, however, this subjectivist conception of value is not an obligatory feature of CDT. It is equally possible to invoke a completely objective notion of value. All that is necessary is that this value should be capable of being measured at least on an interval scale; this is all that is needed to allow the probabilistic notion of the *expected value* of an action to make sense.

As a matter of fact, I favour a version of decision theory that uses an objective notion of value. In fact, I think that what we need is precisely the notion of *choiceworthiness* that I mentioned earlier, which I would interpret as an objective notion of value. So long as choiceworthiness comes in degrees (that is, some courses of action are more choiceworthy than others), and these degrees of choiceworthiness can be measured at least on an interval scale, then we can say that a rational choice is one that maximizes expected choiceworthiness. For our purposes, however, we do not need to worry about exactly what notion of value features in the correct account of rational decision. I shall continue to write as though the relevant notion is this notion of choiceworthiness; but for the purposes of our present discussion, we need only assume that any account of rational decision will use some such notion of value.

For our purposes, it is the following two features of CDT that are particularly crucial. The first of these features consists in the fact that CDT operates with a fundamental distinction between *states* and *acts*. Broadly speaking, we can think of both acts and states as propositions: acts are propositions to the effect that the relevant agent does something of such-and-such an act-type at such-and-such a time, while states can be propositions about more or less anything other than how the relevant agent acts at that time. The difference between states and acts is this: states are completely beyond the agent's control, whereas acts are within the agent's control. It is for this reason that some philosophers – including John Broome (1991, 22) – call these states “states of nature” (although other philosophers – such as David Lewis (1981) – call them “causal dependency hypotheses”). Following Gibbard and Harper (1978), many advocates of CDT take these states of nature to consist of

conjunctions of “non-backtracking” subjunctive conditionals, where each of these conditionals has the form ‘If I did act  $A_n$ , outcome  $O_m$  would result’. For our purposes, however, it does not matter whether or not these states of nature consist of such conjunctions of subjunctive conditionals; the most important fact about them is just that they are completely beyond the relevant agent’s control.

The second feature of CDT that is particularly crucial for our purposes is that, together, an act  $A_i$  and a state  $S_j$  must determine a precise value (or a precise utility, in the subjectivist versions of CDT) – intuitively, the value that the act  $A_i$  would have if it were undertaken in that state  $S_j$ . Nothing else besides this one act  $A_i$  and this one state  $S_j$  is needed to determine this value: in particular, there are no other acts available to the agent that are relevant in any way to determining the value that the act  $A_i$  would have if it were performed in the state  $S_j$ . So the actual value of the act  $A_i$  cannot depend on *anything else* that is within the agent’s control at all. It depends only on factors that are completely beyond the agent’s control; and all of the factors beyond the agent’s control on which the value of this act  $A_i$  depends are included in the state  $S_j$ .

If we just focus on acts of the ordinary kind – acts of the sort that we talk about in everyday life, or of the sort that agents ordinarily make the objects of their choices or intentions – we may well wonder whether it is possible for acts to play this role. Consider for example an act of this ordinary kind, such as *buying a plane ticket to Boston*. The value of this act depends on a great many other things that one does – whether one catches the plane or not, how one behaves while one is in Boston, and so on. So how can it make any sense for CDT to assume that the precise value of every act depends solely on factors that are totally beyond one’s control?

We need to remember here that some acts are more *specific* than others. For example, the act of walking from Nantgwynant to the top of Snowdon is a more specific act than the act of walking somewhere in Wales: it is necessary that anyone who does the first act also does the second, but not *vice versa*. Similarly the acts that are, intuitively, proper parts of other acts – as walking from Nantgwynant to the Gladstone Rock is part of walking all the way from Nantgwynant to the top of Snowdon – are also less specific than the acts of which they are proper parts, since here again, it is possible to do the former act without doing the latter, but not *vice versa*.

I think that we must conclude that the acts that CDT is focusing on are in effect *extraordinarily specific* acts – acts that already include everything that is within the agent’s control that is relevant to determining how good or valuable they are. In the overwhelming majority of cases, the factors that are relevant to determining the value of an act include *both* the end that is achieved *and* the means that are used in order to achieve that end. (The only exceptions to this general rule are cases in which all possible means to a certain end are completely trivial and on a par, so that it makes no difference to the value of one’s act which of these means is used in order to achieve the end.)

In this way, then, CDT cannot model the kind of decision process that proceeds piecemeal, by first deciding on what end to pursue, and then deciding on what means to use to achieve the end. Instead, it can at best serve only as a way of identifying the *complete packages of means and ends* that could be the *total upshot* of an ideally rational process of decision-making. This is why CDT has nothing to say about instrumental rationality. An account of instrumental rationality would be an account of the rational processes of *piecemeal* decision-making, by means of which one could ideally end up taking one of the total courses of action that would be recommended by CDT. But CDT itself does not inquire into what these rational processes of piecemeal decision-making might be.

At least in outline, this point has been recognized by the theorists who have studied CDT. Thus, J. M. Joyce (1999, 70-77) distinguishes “grand-world decisions” from “small-world decisions”. A “grand-world decision” is based on a specification of the relevant acts and states that is so detailed that it captures everything that has any relevance to the value of the acts in question. So, in this “grand-world” specification of acts and states, every act-state pair has a precise value that is independent of the truth or falsity of every other proposition whatsoever. That is, if  $A$  is an act and  $S$  is a state in this “grand-world” specification, then there is no other proposition  $X$  such that  $(A \ \& \ S \ \& \ X)$  is better than  $(A \ \& \ S \ \& \ \neg X)$  in any way. A “small-world” decision, on the other hand, would be based on a less detailed specification of the relevant acts and states – a specification that does not capture absolutely everything that is relevant to the value of these acts.

Traditional CDT is really only designed to be applied to such grand-world specifications of choice situations. But as Joyce (1999, 73-5) has argued, it is doubtful whether any human agent could ever actually make a grand-world decision of this kind. Instead, as I have indicated, we make our

decisions in a more piecemeal fashion; and even though our piecemeal decisions add up to a more detailed and better-informed plan than any of our individual decisions, they will never add up to a full-blown grand-world decision, since we will never be capable of thinking clearly about the huge number of extraordinarily detailed acts and states that such a grand-world decision would have to be based on. So the problem for any theorist who is broadly sympathetic to CDT is to tell some story about how it can be rational to make such “small-world” decisions if it is the “grand-world” decision that ultimately determines which decisions it is ideally rational to make.

**4b.** According to Joyce (1999, 74), the solution to this problem will involve explaining how it is possible for someone who makes a small-world decision to be “*justified* in thinking that her evaluations of the merits of the small-world actions in [the small-world specification of the choice situation] agree with the evaluations that she would give those same acts if she viewed them from the perspective of the grand-world [specification of the choice situation]”.

In my view, however, this is a misdiagnosis of the problem of reconciling small-world decisions and grand-world decisions. To solve the problem, it is not enough to ensure that the evaluation of each “*small-world act*” will be the same whether one evaluates it from the standpoint of a finely individuated specification of the relevant states, or from the standpoint of a more coarsely individuated specification. What matters is not how the agent evaluates each of these small-world acts, taking them one by one. What matters is whether the overall result of the *whole series* of rational piecemeal decisions that the agent makes is at least roughly the same overall course of action as the result of the grand-world decision.

As I shall now explain, Joyce’s solution to the problem does not ensure that the overall result of a series of rational piecemeal decisions is even roughly the same course of action as the result of the grand-world decision. On the contrary, his solution can easily allow cases in which the result of a series of rational piecemeal decisions is a quite different course of action from the result of the “grand-world” decision. As I shall argue, it can easily happen that in one small-world decision – say, the decision between  $A_1$  and  $\neg A_1$  – Joyce’s theory will favour  $A_1$  over  $\neg A_1$ , and in a second small-

world decision – say, the decision between  $A_2$  and  $\neg A_2$  – this theory favours  $A_2$  over  $\neg A_2$ , while in the grand-world decision, this theory favours  $(A_1 \ \& \ \neg A_2)$  over  $(A_1 \ \& \ A_2)$ .

As Joyce (1999, 176) explains, the core of his solution to this problem is a general notion of the “causal value” of propositions that is in his terms “partition-independent”. With such a “partition-independent” notion of causal value, the value of a proposition can always be identified with the probability-weighted sum of the values of the various ways in which that proposition might come true – *regardless* of how exactly these various different “ways in the proposition might come true” are individuated or specified, and in particular regardless of how finely or coarsely they are individuated or specified. In Joyce’s view, it is the great advantage of *evidential* decision theory (EDT) that its conception of the value of propositions is “partition-independent” in this way. So he seeks to devise a conception of “causal value” that has the same kind of partition-independence as EDT’s conception of value.

Strictly speaking, Joyce’s (1999, 178) notion of “causal value” is a two-place relation,  $V(X \setminus Y)$  – intuitively, the value of  $X$  as brought about by  $Y$ . However, the only role that the proposition  $Y$  plays in this value  $V(X \setminus Y)$  is to determine that the probability function involved in defining this value is a special “causal probability” that measures the thinker’s estimate of  $Y$ ’s causal powers to bring about the truth of each of the relevant propositions. Otherwise, this notion of causal value is basically the same as that of *evidential decision theory* (EDT). Each of the possible worlds  $W$  in which  $X$  is true – that is, in effect, each of the maximally specific ways in which  $X$  could come true – has a definite value. Then the causal value of  $X$  is the sum of the values of each of these possible worlds  $W$ , weighted by the quotient of the probability of  $(X \ \& \ W)$  divided by the probability of  $X$  itself. But as almost all probability theorists agree, this quotient of probabilities (if it exists) is the same as the *conditional* probability of  $W$  given  $X$ . So according to Joyce’s conception of causal value, it seems that the causal value of a proposition  $X$  – including a proposition about one’s own acts – is a matter of the degree to which this proposition  $X$  is *evidence* (according to the relevant “causal” probabilities) for each of these possible worlds, along with the relevant value of that world. In short, it is a matter of whether  $X$  is *good news*. This is exactly the same as EDT’s conception of the

evidentially expected value of an act – the only difference being that Joyce’s conception uses “causal” probabilities (instead of the epistemic or doxastic probabilities that are favoured by EDT).

However, it seems that this conception of causal value does not really solve the problem. For example, suppose that out of the grand-world actions,  $(A_1 \ \& \ \neg A_2)$  is ranked as the best,  $(A_1 \ \& \ A_2)$  as second-best,  $(\neg A_1 \ \& \ A_2)$  as second-worst, and  $(\neg A_1 \ \& \ \neg A_2)$  as worst. It might also happen that on the supposition that one does the small-world action  $\neg A_2$ , it is conditionally probable that one will also do  $\neg A_1$ , so that the worst of these grand-world actions  $(\neg A_1 \ \& \ \neg A_2)$  will result. Then  $A_2$  is better news than  $\neg A_2$ , and so Joyce’s theory would rank  $A_2$  as better than  $\neg A_2$ , even though it simultaneously ranks  $(A_1 \ \& \ \neg A_2)$  as better than  $(A_1 \ \& \ A_2)$ .

In this way, then, Joyce’s conception of causal value – together with his assumption that all rational decisions, including both grand-world decisions and small-world decisions, maximize this sort of causal value – seems not to harmonize the small-world decisions and the grand-world decisions in a satisfactory way. Moreover, it is clear that EDT suffers from precisely the same defect. Neither Joyce’s theory nor EDT guarantees that the collective result of a series of rational piecemeal small-world decisions will even roughly coincide with any of the courses of action that would be selected by a rational grand-world decision. We need an alternative solution to the problem.

5. In Section 3, we saw that rational practical reasoning involves some kind of estimate of (a) the *availability*, and (b) the *value* or *choiceworthiness*, of various options. Then in Section 4, we saw that it also requires (c) some kind of *integration* of the various “small-world” piecemeal decisions that one makes so that they collectively lead to broadly the same course of action as a rational “grand-world” decision. Presumably, this sort of integration will have to involve some estimate of both the availability and the value or choiceworthiness of the overall upshot of these piecemeal decisions.

One promising way for a theory of practical rationality to require this sort of integration is for it to impose a general constraint on the agent’s whole set of intentions. Following Michael Bratman (1987), we could say that the total set of intentions that an agent has at any one time forms an overall “plan”. Then our theory of practical rationality could require that at every time, the agent’s total set of

intentions should make it (in some appropriate way) *likely* that the plan formed by these intentions is (in some appropriate way) both (a) available and (b) valuable or choiceworthy.

In Section 1, I proposed that believing that an option is “available” involves having a high conditional probability – in the circumstances amounting for all practical purposes to conditional certainty – that if one intends the option, one will execute one’s intention and act accordingly; and in Section 2, I suggested that for an intention to be rational, it must also be rational for the agent to believe the intended option to be available in this way. This requirement applies to every single intention taken by itself. I now propose that we should supplement this requirement on each particular intention with a parallel general constraint on the agent’s whole set of intentions.

One suggestion that might be made is that to be rational, one’s entire set of intentions must be such that one attaches an equally high conditional probability to the proposition that one will execute *all* those intentions, on the assumption that one has those intentions. But this would be an extremely strong requirement. Indeed, it seems more plausible that if I am rational, I will be much *less* confident that I will execute *all* of my current intentions than that I will execute my intention to get to the top of Snowdon on the Watkin path. So it would seem more plausible to require merely that if one is rational, one’s intentions should *not* be such that one has a high conditional probability that if one has precisely those intentions, one will *not* execute all of one’s intentions.<sup>13</sup> Even though this requirement is weaker than the extremely strong requirement that I have just considered, it seems enough to yield Bratman’s “strong consistency” condition. If the contents of one’s intentions and of all the beliefs that one holds with certainty form an inconsistent set of propositions, then it will be presumably be rational for one to be certain that either (i) some of these beliefs that one holds with certainty are false,

---

13 As Cian Dorr has reminded me, it would be implausible to impose the analogous condition on beliefs. As the preface paradox shows, even if one is a perfectly rational believer, it can be rational for one to have a high conditional confidence that not all of one’s beliefs are true, conditionally on the assumption that one holds those beliefs. Still, I do not think that it is implausible to impose this requirement on intentions. The solution to the preface paradox will have somehow to invoke the idea of “degrees of belief”; but while there are degrees of belief, there are no degrees of intention. So it is not surprising that intentions are subject to more demanding requirements than beliefs.

or (ii) one will not execute all of one's intentions. If the reason for being in this position is not that this set of beliefs is itself inconsistent (which would presumably also be irrational), then one will violate the general constraint on one's intentions that I have just proposed.

With respect to the dimension of value or choiceworthiness, it is tempting to suggest that a rational agent will have a set of intentions that in some way maximizes expected value. However, we might wonder whether it is the set of *intentions* or the set of intended *courses of action* that must maximize expected value. On the one hand, it might seem more plausible to require that the intentions must maximize expected value. There are cases in which it is possible to intend a course of action that one knows perfectly well one will in fact take even without intending to. If having the intention will not help to improve how one acts, then it need not be irrational to lack the intention even if the course of action in question is part of an overall course of action that *does* maximize expected value. In general, just because something is a good thing to *do*, it does not follow that it must also be a good thing to *intend* to do; having the intention may not help.

On the other hand, it might seem that if our theory requires that our *intentions* should maximize expected value, we will be committed to saying that the rational agent must intend to drink the toxin in Gregory Kavka's (1983) "toxin puzzle" case. (This is the case in which an eccentric billionaire will give you £1 million at midnight tonight if you then *intend* to drink the toxin tomorrow morning, but is quite indifferent to whether or not you *actually* drink the toxin tomorrow morning; so in this case, *having* the intention is beneficial, but *executing* the intention is not.) Yet intuitively it seems doubtful whether the advantages of intending to drink the toxin are enough to make it rational to have this intention.

In fact, however, it is possible to finesse these tricky issues. Our theory should require that the agent's intentions should maximize expected value in the following special sense. To assess a given set of intentions, the relevant probabilities to use are the *conditional* probabilities – the probabilities of the various relevant propositions *conditional on one's having that set of intentions*. However, the propositions whose probability is in question are propositions about the value of the *course of action* that one would take if one were to execute those intentions.

This approach will not require the agent to intend a good state of affairs if it is clear that having the intention will not help to bring about that state of affairs; a set of intentions that includes useless (or counter-productive) intentions of this kind will not have a higher “expected value” than a set that lacks them. It is also clear that this approach will not require intending to drink the toxin, since although intending to drink the toxin is evidence that one is in a wonderful fortunate situation – namely, a situation in which the eccentric billionaire will give one £1 million – it is not evidence that *acting* on one’s intentions is good or valuable in any way. In this way, this approach succeeds in finessing both of these worries.

However, we still need to clarify how exactly we are thinking of the kind of “expected value” that is in question here. It seems plausible to me that CDT is right that what we may call the “objective rightness” of the available acts depends on how all the most specific and most detailed of these acts compare with each other, in all the possible worlds in which all the facts that are beyond one’s control are held fixed. Out of these maximally detailed and specific acts, an act is objectively right if and only if it is optimally choiceworthy (that is, no less choiceworthy than any of the alternative acts); out of the less detailed and less specific acts, an act is objectively right if and only if it is *part* of a more specific act that is objectively right. In this way, CDT is right about what determines the degree of objective rightness or wrongness that each of these available acts has. But as we have seen, actually *considering* each of these highly specific acts, and attaching a definite probability to each state of nature that assigns a definite degree of choiceworthiness to each of these highly specific acts, seems a feat that is beyond the capacities of human reasoners. So although CDT is right about objective rightness and wrongness, it seems wrong about *practical rationality*.<sup>14</sup>

---

14 This distinction between objective rightness and practical rationality may help to explain why the relevant probabilities for rational decision-making are *conditional probabilities*, rather than the probabilities of *subjunctive conditionals*. Subjunctive conditionals are indeed of crucial importance in determining the degrees of objective rightness or choiceworthiness that the available courses of action have, since these degrees of choiceworthiness depend on how the agent’s actual course of action compares with various *possible* but *non-actual* alternatives. But rational decision-making involves evaluating the various available plans as ideas about what one will *actually* do (not as ideas about how things are in non-actual possible worlds). So evaluating these

However, even if ordinary agents cannot easily think about the full complexity of all the facts that determine the degree to which each of these acts is objectively right or wrong, they *can* think in more general terms about the choiceworthiness of the conduct that they are embarking on. That is, they can entertain propositions of the form, ‘Out of all the currently available courses of action, the course of action that I will actually take will be good or choiceworthy to degree  $d$ ’. I shall call propositions of this form “value-specifying propositions”. I propose that we should use propositions of this sort to define the relevant notion of the “expected value” of a set of intentions, together with the conditional probabilities that I mentioned (that is, the probabilities of the various relevant propositions conditional on one’s having this set of intentions).

Once we have these conditional probabilities, and these value-specifying propositions, we can define a notion of the expected value of this set of intentions in the standard way. Given a partition of such propositions – that is, a set of such value-specifying propositions such that one is rationally certain that exactly one of these propositions is true – we can define the relevant “expected value” of this set of intentions as the weighted sum of the degrees of value that are specified by these value-specifying propositions, weighting each degree of value by the relevant conditional probability of the relevant value-specifying proposition. To put it roughly but more intuitively, a rational agent will have a set of intentions that collectively constitutes good evidence for believing that acting on these intentions would be a suitably good thing to do.<sup>15</sup>

---

plans involves exploring them as hypotheses or suppositions about the *actual* world – and as Joyce (1999, Chap. 6) explains, this is the distinctive function of conditional probabilities.

15 This appeal to conditional probabilities – probabilities that are conditional on the assumption of one’s having the intentions in question – raises the question whether this approach coincides with evidential decision theory (EDT) in such cases as the notorious “Newcomb problem”. The answer depends, in my view, on how exactly we measure the *value* or *choiceworthiness* of courses of action. As I hope to explain in future (in my paper “Gandalf’s Solution to the Newcomb Problem”, in progress), on the intuitively most plausible ways of measuring the value or choiceworthiness of courses of action, we should agree with CDT (*not* EDT) about the Newcomb problem, even if the appropriate probabilities are indeed conditional probabilities, of the kind that are invoked by EDT.

Some philosophers may doubt whether it is any more possible to adjust one's intentions so that they maximize expected value in this way than to make a "grand-world" decision of the sort that in the last section I described as beyond the capacity of ordinary human reasoners. But I am not saying that ordinary rational agents must explicitly think about this constraint on rational intentions, or that they should consciously aim to conform to it. Indeed, ordinary agents can conform to this constraint without using the concept of probability at all. In adjusting their intentions in such a way as to conform to this constraint, these agents are responding directly to the overall profile of the relevant conditional probabilities of these propositions themselves. These conditional probabilities can be identified either with rational conditional beliefs,<sup>16</sup> each held with the appropriate degree of confidence – or perhaps with whatever facts about the agents' evidence make it rational for them to have this conditional belief with that degree of confidence. I do not see any decisive reason why it should not be possible for ordinary agents to adjust their intentions in such a way that they conform to this constraint. (Moreover, the most that is required for rational requirements is *possibility*: there is no reason to assume that being rational must be easy!)

The picture that emerges from this proposal, then, is the following: as the evidence comes in, the rational agent will continually adjust her intentions to her evidence, in such a way that at every time, her complete set of intentions, together with this evidence, makes it rational for her to have the appropriate beliefs about the availability and value of the intended course of action. Specifically, these intentions, along with the evidence, must meet the following conditions: (i) they must make it rational for the agent to have a high level of confidence, concerning each of these intentions, that she will execute that intention, and they must not make it rational for her to have a high level of confidence that she will not execute all of these intentions; and (ii) they must make it rational for the agent to have a degree of conditional confidence in each of the "value-specifying" propositions, conditionally on the assumption that she has that set of intentions, so that the set of intentions maximizes "expected value" in the way that I have described.

---

<sup>16</sup> I do not mean a belief in a conditional proposition; I mean an intrinsically conditional belief, of the sort that has been defended by Edgington (1995).

As I shall now argue, this proposal can capture the intuitive features of instrumental rationality that I identified in Section 2. The main requirement of instrumental rationality that I defended in Section 2 was the principle that it is irrational simultaneously to intend an end, to believe that a certain set of possible means are the optimal means to that end, and yet to intend none of those means to the end; or at least, in any such case it is irrational to persist with not intending any of these means if one also rationally believes that one will not achieve the end in an optimal way unless one now decides on some appropriate means. As I shall now explain, there is a reasonable interpretation of this principle on which it can be derived from the proposal that I have just made about what it is for our “small-world” piecemeal intentions to be rational.

Suppose that one did have the combination of attitudes that this principle outlaws as irrational. That is, suppose that one intends an end, and believes that one will not achieve this end in an optimal way unless one now intends to take one of a certain set of means, but also intends none of these means. Plainly, this combination of attitudes could be irrational if it is intrinsically irrational to intend this end. It could also be irrational if it is intrinsically irrational to hold the belief that one will not achieve the end in an optimal way unless one now decides to take one of the means in question. But suppose that neither this intention nor this belief is intrinsically irrational. Why then is there anything irrational about this combination of attitudes?

To see what is irrational in these cases, we shall have to reinterpret my earlier talk of “believing certain means to be optimal” in a certain way. Specifically, we shall have to reinterpret it so that one believes each member of a certain set of courses of action to be an optimal means to a certain end if and only if each intention to achieve the end by one of those means has greater expected value than either (i) the intention to achieve the end by any *alternative* means, or (ii) the bare intention to achieve the end without intending to use *any* particular means in order to do so. Once we reinterpret this belief in this way, it is clear that any agents who have this belief, intend the end, but do not intend any of the courses of action that they believe to be optimal means to the end, cannot have a set of intentions that maximizes expected value in the way that I have just described. In this way, we can capture the fact that this combination of attitudes counts as instrumentally irrational.

This conception of rational intentions can also clarify the two doubtful points that I mentioned at the end of Section 2. First, according to this conception, it is not necessary that every single rational intention should be for some course of action that one rationally believes to be optimal; it is only required that each intention should belong to a whole set of intentions that maximizes expected value in the appropriate way. If one is rational, one need not evaluate every single thing that one intends separately; one may evaluate many of the smaller things that one intends as parts of larger packages of intentions. Secondly, it is clear that by appealing to the expected value of one's set of intentions, this conception only needs to suppose that one has *partial* conditional beliefs (or *degrees of confidence*) in various value-specifying propositions about the value of the course of action that one is embarking on. It does not require that one should have an outright, all-or-nothing belief in any proposition that ascribes optimality to any available course of action. So, even if one does not know which of the available courses of action are optimal and which are not, it is possible to be guided by rational partial conditional beliefs of this sort.

6. In this essay, I have identified three aspects of practical rationality. To sum them up roughly, they are the following:

- (a) Adjusting one's intentions to one's beliefs or expectations about what courses of action are *available*.
- (b) Adjusting one's intentions to one's beliefs or expectations of the *value* or *choiceworthiness* of the course of action that will result from those intentions.
- (c) *Integrating* the various different intentions that one forms, in the course of piecemeal "small-worlds" practical reasoning, into a rational overall plan.

I have suggested that what is characteristic of instrumental reasoning is especially this third element (c). We engage in instrumental reasoning only because we reason in a piecemeal fashion, deciding first on the end and only later making up our minds about which means to use to achieve the end. In the proposal that I made in the previous section, this third element is reflected in the requirement that

the agent's whole *set* of intentions must collectively form a plan that has certain rational expectations of availability and value. Specifically, these intentions (a) must make it rational for the agent to believe, concerning each of these intentions, that she will execute that intention, and not make it rational for her to believe that she will not execute all of those intentions; and (b) this set of intentions must also maximize expected value in the way that I have described.

However, it is fair to point out that this third element has no *necessary* connection with the process of reasoning from ends to means. It is also possible, at least in principle, to reason in the opposite direction, from means to ends. One might decide that one needs a walk and only later decide on the destination that one's walking will be aimed at. It may be that the commonest form that the process of integrating one's different intentions into a rational overall plan can take is when we reason from ends to means; but it is not the only possible form that this process can take.

In these cases, some philosophers will be tempted to reply that one's real end is just to *have a walk* and the action of *walking to Chilham*, say, is one's means to that end. But given the sense in which I am using the term 'end', this reply would be mistaken. As I am using the term, one's intention for the "end" need not be the intention that *motivates* one's intention for the means; nor need it be the intention that one takes to *justify* or reveal what is *worthwhile* about one's intention for the means. One's intention for the end is simply the intention that (as I put it) "controls" or "guides" one's execution of the intention for the means. In this case, the intention to get to Chilham serves in the most literal sense to "guide" one's steps: one puts one foot in front of the other in precisely such a way as to take one to Chilham (rather than, say, to Wye, or anywhere else). In this way, the intention to get to Chilham guides the execution of one's intention to go for a walk. So going to Chilham counts as the intended end, while walking is one's means to that end – even if it was the intention to go for a walk that motivated one's intention to go to Chilham, and even though going to Chilham only seemed a good or worthwhile thing to do because it involved going for a walk. So there is no reason to deny that the process of filling out one's schematic intentions and integrating them into a rational overall plan may sometimes take the form of reasoning from means to ends, as well as the more common form of reasoning from ends to means.

Nonetheless, our investigation of these three aspects of practical rationality has led us to a better understanding of the features of rational instrumental reasoning that I sketched in Sections 1 and 2. In that sense, we have made progress towards understanding the phenomena that philosophers have discussed under the rubric of “instrumental rationality”.<sup>17</sup>

---

<sup>17</sup> I am grateful to audiences at Harvard University, Brown University, and the University of Oxford, for helpful comments on earlier drafts. The final revisions were carried out during my tenure of a Research Fellowship from the Leverhulme Trust, to whom I should also like to express my gratitude.

## References

- Bratman, Michael E. (1987). *Intentions, Plans, and Practical Reason*. Cambridge, Massachusetts: Harvard University Press.
- Broome, John (1991). *Weighing Goods* (Oxford: Blackwell).
- (1999). “Normative requirements”, *Ratio* 12 (4): 348-419. DOI: 10.1111/1467-9329.00101
- (2005). “Have we reason to do as rationality requires? A comment on Raz”, *Journal of Ethics and Social Philosophy* Symposium I (December): 1-8. <http://www.jesp.org/>
- Dreier, James (1997). “Humean doubts about the practical justification of morality”, in Garrett Cullity and Berys Gaut, eds., *Ethics and Practical Reason* (Oxford: Clarendon Press): 81-100.
- Edgington, Dorothy (1995). “On Conditionals”, *Mind* 104: 235-329
- Gibbard, Allan, and Harper, William (1978). “Counterfactuals and Two Kinds of Expected Utility”, in C. A. Hooker et al., eds., *Foundations and Applications of Decision Theory* (Dordrecht: Reidel), vol. 1: 125–62.
- Jeffrey, Richard (1983). *The Logic of Decision*, 2nd edition (Chicago: University of Chicago Press).
- Joyce, J. M. (1999). *Foundations of Causal Decision Theory* (Cambridge: Cambridge University Press).
- Kant, Immanuel (1785). *Grundlegung zur Metaphysik der Sitten* (Riga: Hartknoch). Cited by the pagination of the Royal Prussian (later German) Academy edition (Berlin: 1900–), vol. 4.
- Kavka, Gregory (1983). “The Toxin Puzzle”, *Analysis* 43, 33–36.
- Kolodny, Niko (2005). “Why be rational?”, *Mind* 114 (3): 509-63. DOI: 10.1093/mind/fzi509
- Lewis, David (1981) “Causal Decision Theory”, *Australasian Journal of Philosophy* 59: 5–30. Reprinted in Lewis (1985, 305–337).
- (1985). *Philosophical Papers*, Vol. II (Oxford: Clarendon Press).
- Mele, Al (2000). “Goal-directed action: Teleological explanations, causal theories, and deviance”, *Philosophical Perspectives* 14: 279-300.
- Raz, Joseph (2005). “The myth of instrumental of rationality”, *Journal of Ethics and Social Philosophy* 1 (1): 2-28. <http://www.jesp.org/>
- Schroeder, Mark (2004). “The scope of instrumental reason”, *Philosophical Perspectives* 18: 337-364

- (2009). “Means-end coherence, stringency, and subjective reasons”, *Philosophical Studies* 143: 223-249. DOI: 10.1007/s11098-008-9200-x
- Setiya, Kieran (2007a). “Cognitivism about instrumental reason”, *Ethics* 117 (4): 649-73.  
DOI: 10.1086/518954
- (2007b). *Reasons without Rationalism* (Princeton, New Jersey: Princeton University Press).
- Wedgwood, Ralph (2002). “Practical Reason and Desire”, *Australasian Journal of Philosophy* 80: 345-358.
- (2003). “Choosing Rationally and Choosing Correctly”, in *Weakness of Will and Practical Irrationality*, ed. Sarah Stroud and Christine Tappolet (Oxford: Oxford University Press, 2003): 201-229.
- (2004). “The Metaethicists’ Mistake”, *Philosophical Perspectives* 18 (2004): 405-426.
- (2007). *The Nature of Normativity* (Oxford: Clarendon Press).
- Williams, Bernard (1981). “Internal and external reasons”, in his *Moral Luck* (Cambridge: Cambridge University Press).