Organon F 20 (4) 2013: 484-506

Akrasia and Uncertainty*

RALPH WEDGWOOD

School of Philosophy. University of Southern California
Los Angeles. CA 90089-0451. USA
wedgwood@usc.edu

RECEIVED: 29-12-2012 • ACCEPTED: 16-07-2013

ABSTRACT: According to John Broome, *akrasia* consists in a failure to *intend* to do something that one believes one *ought* to do, and such *akrasia* is necessarily irrational. In fact, however, failing to intend something that one believes one ought to do is only guaranteed to be irrational if one is *certain* of a *maximally detailed* proposition about what one ought to do; if one is uncertain about any part of the full story about what one ought to do, it could be perfectly rational not to intend to do something that one believes one ought to do. This paper seeks to remedy this problem, by proposing an anti-*akrasia* principle that covers cases of uncertainty (as well as cases of such complete certainty). It is argued that this principle is in effect the fundamental principle of practical rationality.

KEYWORDS: Act individuation – *akrasia* – John Broome – decision theory – practical rationality – probability – uncertainty.

1. The irrationality of *akrasia*

According to Socrates and Aristotle – at least as I shall interpret them here – *akrasia* (if it is possible at all) would involve *voluntarily* doing something that one *knows* one *ought*, *all-things-considered*, *not* to do.[1] It is widely

* Earlier versions of this paper were presented as talks at Auburn University and at Princeton University. I am grateful to both audiences for extremely helpful comments.

[1] For Socrates' investigations of *akrasia*, see especially Plato's *Protagoras* (351a–358d); for Aristotle's discussion, see *Nicomachean Ethics* VII.1–10.

agreed that if such *akrasia* is possible, it is irrational – indeed, it is a paradigmatic form of irrationality. To be *akratic* in this way at a given time *t*, by voluntarily doing something at *t* that one simultaneously knows one ought not to do, is incompatible with being fully rational at that time *t*. As I shall use the term here, to say that you are at a given time *t* "rationally required" to φ is just to say that it is *necessary* for you to φ if you are to be fully rational at *t*.[2] So it seems that every agent is rationally required not to be *akratic* at any time.

In this essay, I shall not investigate the question that Socrates and Aristotle puzzled over, of how such *akrasia* is possible. Instead, I shall focus on a more basic question: assuming that we are rationally required not to be *akratic*, what is the most precise account of the general principle that underlies this rational requirement?

I shall start my exploration of this question by considering the formulation of this "anti-*akrasia*" principle that is given by John Broome. We shall see that Broome's principle is unsound unless it is restricted to a narrow range of cases – specifically, to cases where there is an extremely *fine-grained* way of carving up the available options or courses of action, such the agent in question is for all practical purposes *certain* about which of these fine-grained options she ought to do.

Nonetheless, if it is restricted in this way, Broome's principle seems to be sound. This raises the question of how we can generalize this restricted version of this principle so that it covers a wider range of cases, where the agent is uncertain about which of these fine-grained options she ought to do. The last three sections of the paper are devoted to this question.

2. Broome's principle

When Socrates undertook his investigations of *akrasia*, he assumed that if it existed at all, it would consist in voluntarily doing what one *knows* one ought not to do. But in fact, it seems plausible that even if you merely *believed*, and did not know, that φ-ing was something that you all-things-

[2] Some philosophers – most notably, John Broome (2013) – understand the phrase 'rational requirement' rather differently. In my view, however, this is simply a terminological issue: the phrase 'rational requirement' is a semi-technical term, which is not in regular use by ordinary citizens; so it is quite legitimate for me simply to stipulate how I shall use the term here.

considered ought not to do, you would still be being *akratic* if in spite of this belief, you were voluntarily to φ. In such cases, there is a kind of conflict or incoherence between your beliefs and your will: your beliefs in some sense tell you not to φ, while your will voluntarily embraces φ-ing. Such incoherence or conflict between your beliefs and your will seems to be irrational. Rationality requires us to avoid such conflicts.

John Broome accepts this point that rationality requires a kind of coherence between one's beliefs and one's will. He has attempted to give a more precise account of exactly what this coherence consists in (see Broome 2013, Section 6.5). In order to articulate this account as precisely as possible, he makes use of several technical or semi-technical terms (which he first defines in careful detail). Since many aspects of his account are not central to the issues that I shall explore in this essay, I shall take the liberty of rephrasing his account in slightly more ordinary terminology. This rephrasing will not change the meaning of Broome's principle in any respects that are relevant to the purposes that I am pursuing here – but it should *not* be taken as a completely accurate presentation of Broome's view for all purposes whatsoever.

For our purposes, then, we may interpret Broome's principle as equivalent to this:

> Rationality requires of you that:
> If
> (1) You believe at t that you yourself ought to φ, and
> (2) You believe at t that, if at that time you intended to φ, then because of that, you would indeed φ, and
> (3) You believe at t that, if at that time you did not intend to φ, then because of that, you would not φ,
> Then
> (4) You must intend at t to φ.

The purpose of conditions (2) and (3) of this principle is effectively just to narrow down the scope of this requirement to cases in which, if the beliefs mentioned in these conditions are correct, you will avoid voluntarily failing to φ if and only if you *intend* at t to φ. I shall not worry about this aspect of Broome's principle here. I shall simply focus on cases where these two conditions (2) and (3) are met. The main focus of my discussion here will be on the relation between condition (1), believing at t that you ought to φ, and condition (4), intending at t to φ.

Broome's claims about this principle have already been widely debated (see, for example, Kolodny 2005). But it seems plausible, at least to me and to many others, that there is at least *one* interpretation of the principle such that whenever you violate Broome's principle, on this interpretation of what it means, you are being *akratic* – and so irrational. The main task of this essay is to work out what exactly this interpretation of the principle is.

One of the main issues that arise about the principle concerns the interpretation of 'ought'. In fact, the word 'ought' in English and its equivalents in other languages seem to be systematically polysemous, and capable of expressing a range of different concepts in different contexts. In effect, 'ought' has many different senses in different contexts. So, one of the central questions that we have to address is this: Which senses of 'ought' will make all instances of Broome's principle true?

Suppose that Broome's principle is true in all instances for *more than one* sense of 'ought'. Now, unless one of these senses of 'ought' analytically implies the other, it would surely be possible, at least in principle, for you to be in a case in which you are rationally required to believe that in one of these senses you "ought" to φ, and also required to believe that in another of these senses you "ought not" to φ. (For example, suppose that an oracle whose pronouncements have the most extraordinary track record for reliability announces that you are in a case where you "ought" in the first sense to φ, but "ought" in the second sense not to φ. Then it seems that you could be rationally required to have both beliefs.)

In that case, however, if Broome's principle were true in all instances for both of these senses of 'ought', you would be simultaneously rationally required to intend to φ and rationally required to intend *not* to φ. But having contradictory intentions of this sort also seems paradigmatically irrational – not something that can result from one's complying with rational requirements.

For this reason, I shall assume from now on that there is just *one* sense of 'ought' for which Broome's principle is true in all instances. I shall return later on to the question of what exactly that sense of 'ought' is. For the time being, I shall just try to read the principle sympathetically and charitably – that is, in effect, to read the principle as involving that sense of 'ought', whatever it is, that makes it most plausible that the principle is true in all instances.

3. Two issues in the philosophy of belief and intention

There are two extremely well-known issues in the philosophy of belief and intention, which seem obviously relevant to the evaluation of Broome's principle.

i. Beliefs come in *degrees*; we believe some propositions *more strongly* or *more confidently* than others.
ii. The *acts* or *options* that an agent can intend to perform can be *individuated* in different ways – sometimes more *finely* and sometimes more *coarsely*.

To illustrate the first issue, we may note that it seems that I believe the propositions that I exist, and that 1 + 1 = 2, with more confidence than the proposition that Dushanbe is the capital of Tajikistan. So, when Broome's principle refers to what "you believe" (as it does in each of its first three clauses (1), (2), and (3)), we need to know: What degree of confidence must you have in the relevant proposition for it to be true in this context to say that you "believe" the proposition?

To illustrate the second issue, we need to remember that some kinds of acts are very general and unspecific, like *moving one's hands*, while others are much more specific, like *signing a cheque to pay November's phone bill*. In this example, the first kind of act is less specific than the second kind, because it is necessary that whenever one performs an act of the second kind, one also performs an act of the first kind, but not *vice versa*. So, again, we need to know, when Broome's principle uses the schematic letter 'φ', can this letter 'φ' take the place of any act-description, or can it only take the place of an act-description that is at a certain level of generality or specificity?

Broome's principle seems most compelling when the following two conditions are met:

i. The relevant *beliefs* (referred to in clauses (1), (2), and (3) of the principle) are beliefs held with *maximum confidence*.
ii. The act of *your* φ*-ing* (referred to in every clause of the principle) is individuated extremely *finely*, so that it is a highly *specific* act, capturing *everything* of importance in the relevant situation.

When these conditions are met, you are *totally convinced* of a proposition that – if true – gives the *whole truth* about what you ought to do in

your situation. If you voluntarily act contrary to a conviction of this sort, you are surely being irrational in some way.

However, this is a severely restricted version of Broome's principle. This restricted version of the principle says nothing about the cases where you cannot be *rationally certain* of any such highly detailed proposition about what you ought to do. It seems clear that such uncertainty can undoubtedly arise, since facts about what you ought to do seem not to be, as we might put it, *rationally luminous*. That is, it is not in general the case that *whenever* such a fact obtains, it is possible for you rationally to have an attitude of *maximum confidence* in the proposition that that fact obtains. Indeed, a version of the famous "margins for error" argument of Timothy Williamson (2000, 93-106) seems to show that normative facts cannot be rationally luminous in this way. There could be a continuous series of cases, such that in the case at the *beginning* of the series, the only attitudes that it is possible for you rationally to have towards the proposition *that you ought to* φ all involve a high level of confidence in that proposition, while in the case at the *end* of the series, the only attitudes that it is possible for you rationally to have towards that proposition all involve a high level of *disbelief* in that proposition (and so also a high level of confidence in the negation of the proposition). We may also make two further stipulations about this series of cases: first, there are no cases in this series in which it is *both* possible for you rationally to have a high level of confidence in the proposition that you ought to φ and *also* possible for you rationally to have a high level of disbelief in this proposition; and secondly, for every case in the series after the very first case, the range of attitudes that it is possible for you rationally to have in that case differs at most only *very slightly* from the range of attitudes that it is possible for you rationally to have in the immediately preceding case.

Given these stipulations, it follows that there must be some cases, somewhere in the middle of this series, where the only attitudes that it is possible for you rationally to have in the proposition that you ought to φ (if indeed it is possible for you rationally to have *any* attitudes towards that proposition at all) are all *intermediate* levels of confidence, rather than high levels of confidence or high levels of disbelief. In those cases, given classical logic, either the normative proposition that you ought to φ is true, or its negation is true – where this negation is equivalent to the normative proposition that it is *permissible* for you *not* to φ. Either way, then, there is a true normative proposition in which it is not possible for you rationally to

have a high level of confidence. Thus, true normative proposition are not (as I put it) rationally luminous. Cases can arise in which it is impossible for you to be rationally *certain* about what you "ought" (in the relevant sense) to do.

Now, it may be that in at least some cases of this kind, there is some *other* sense of 'ought' such that you *are* rationally certain about what you "ought", in this other sense, to do. Even if that is true, however, it is irrelevant. We are assuming that there is exactly one sense of 'ought' that features in Broome's principle. The cases that we are considering are cases in which one is not certain about what one "ought" to do *in this crucial sense*. It is irrelevant if in some of these cases, there is some other sense of 'ought' such that you are certain of what you "ought" in that other sense to do. That other sense of 'ought', whatever it may be, is not the sense that appears in Broome's principle, and so need not concern us here.

It is crucial to see that this Williamson-inspired argument shows that *no* non-trivial sense of 'ought' is rationally luminous: for every non-trivial sense of 'ought', cases can arise in which a rational agent cannot be certain about what she ought to do. For this reason, any version of Broome's principle that is restricted to cases in which the agent is certain about what she ought to do is, as I have said, a severely limited principle. The significance of this point will emerge in the sequel.

4. Against the unrestricted form of Broome's principle

In this section, I shall consider an *unrestricted* form of Broome's principle. In this unrestricted form of the principle:

i. The "beliefs" referred to in the principle can be held with *any* degree of confidence that is at least as great as some threshold t, where $t < 1$.
ii. The schematic letter 'φ' can stand for *any* act, regardless of whether it is a finely individuated, highly specific act or a coarsely individuated, highly general act instead.

As I shall argue here, this unrestricted form of Broome's principle is open to fatal counterexamples. I shall start by presenting a counterexample in which *none* of the agent's beliefs are held with the maximum level of confidence; this counterexample does not depend on the issue of how finely or coarsely the relevant acts are individuated. Then I shall present a

second counterexample, in which the agent is *certain* about which *coarse-grained* acts she ought to do, and is uncertain only about which *fine-grained* acts she ought to do.

4.1. Counterexample (i) to the unrestricted form of Broome's principle: Uncertainty about all options

The first counterexample to the unrestricted form of Broome's principle is a case in which there are two options available to you, A and B, such that these two options form a *partition* – i.e., you are certain that you will do *one*, and *no more than one*, of these two options.

In this case, although you are *not certain* whether you ought to do A, or ought to do B, you have a *very high* degree of confidence that you ought to do A – a degree of confidence x such that $x \geq t$. In other words, your degree of confidence that you ought to do A is at least as great as the crucial threshold t. Still, in your view, you cannot absolutely rule out the rival hypothesis that you ought to do, not A, but B instead; and so your degree of confidence x in the proposition that you ought to do A is less than certainty – that is, $x < 1$.

Now, let us also assume that in this case, you are *conditionally* certain, given the assumption that it is *not* the case that you ought to do A, that B is not just slightly better than A, but *astronomically* better than A. (For example, perhaps, if it is not the case that you ought to do A, doing A will result in the destruction of the whole world or the like.) On the other hand, you are also conditionally certain, given the assumption that it is the case that you ought to do A, that A is only very *slightly* better than B.

In this case, it seems possible for you to be rational, to have beliefs of this sort, and simultaneously to intend to do not A, but B instead. If that is right, then this case is a clear counterexample to the unrestricted version of Broome's principle. You are perfectly rational, you have a degree of belief above the threshold t that you ought to do A, and yet you do not intend to do A – you intend not to do A, but to do B instead.

Some readers might suspect that we can get round counterexamples of this sort simply by amending the unrestricted form of Broome's principle so that the kind of 'belief' referred to in the principle must consist of beliefs of which the agent is, for all practical purposes, completely *certain*. As we shall see in the next subsection, this suspicion is incorrect: this amendment does not make the principle immune to counterexamples of this kind.

4.2. Counterexample (ii) to the unrestricted form of Broome's principle: Uncertainty about the fine-grained options

Our second counterexample concerns a case – like Frank Jackson's "three drugs" case[3] – where there are *three* fine-grained options available to you: *A*, *B*, and *C*. Again, suppose that these three options form a partition (that is, you are certain that you will do exactly one of these three options). There are also some coarse-grained options, like *doing A or B*, or *doing B or C*, or *not doing A*, and so on. Since *A*, *B*, and *C* form a partition, the coarse-grained option of *not doing C* and the coarse-grained option of *doing A or B* are effectively equivalent.

In this case, suppose that you are certain that *either* you ought to do *A* *or* you ought to do *B*; and you are also certain that you ought *not* to do *C*. However, you are radically *uncertain* about whether the option that you ought to take in this situation is *A* or *B*.

In addition, in this case, you are *conditionally* certain, given the assumption that you ought to do *A*, that doing *B* will be utterly disastrous (it will result in the destruction of the world or the like), and you are also conditionally certain, given the assumption that you ought to do *B*, that doing *A* will be equally disastrous. However, you are *also* certain that doing *C*, though it falls short of being what you strictly ought to do, is not *too* bad: it is far less bad than doing *A* would be if you ought instead to have done *B,* and equally far less bad than doing *B* would be if you ought instead to have done *A*.

In this case, you might be rational, and be certain that you ought not to do *C* (or, equivalently, that you ought to do either *A* or *B*), without intending not to do *C* (or, equivalently, without intending to do *A* or *B*). If that is right, then this case is also a counterexample to the unrestricted form of Broome's principle: you are rational, you are certain that you ought not to do *C,* and yet you do not intend not to do *C*.

In this way, then, the unrestricted form of Broome's principle seems to be faced with fatal counterexamples. This unrestricted form of the principle is unacceptable.

---

[3] See Jackson (1991); another famous case of this sort is the "three mineshafts" case of Parfit (2011, 159), which was inspired by an example of Regan's (1980, 264-265, n. 1).

5. A better way of generalizing the restricted form of Broome's principle

Still, as I commented above, the restricted form of Broome's principle – restricted to cases where the agent is *certain* about which *fine-grained* option she ought to do – seems compelling. It does seem necessary that if an agent is perfectly rational, and is certain of the truth of such a fully specific proposition about what she ought to do, then the agent will intend to do the fine-grained act that she is certain she ought to do.

However, as I argued in Section 3, the restricted form of Broome's principle is severely limited to a narrow range of cases. It seems plausible that if this restricted principle holds in this narrow range of cases, that will be because of some more general truth that explains why it holds in these cases. But what is this more general truth? How can we generalize this restricted form of the principle, in order to cover cases of uncertainty about which fine-grained option the agent ought to do, while avoiding these troublesome counterexamples?

5.1. Generalizing (i): Uncertainty

First, let us just focus on the issue of *uncertainty*. Let us leave aside the issue of option-individuation, for the time being – let us simply assume that we are considering only super-finely individuated options.

It seems clear that in each of the troublesome cases that we have just considered, the rational intention is an intention that *maximizes* some kind of *expectation* of some kind of *value*. We might try to treat the principle that rational intentions maximize the relevant sort of expected value as if it were a completely separate principle from the restricted form of Broome's principle. But it seems as if it would be more promising to unify these principles somehow. This is what I shall try to do here. Specifically, I shall try to find a version of the idea that rationality requires us to have intentions that maximize expected value which implies the restricted form of Broome's principle as a special case.

In defining any notion of expected value, we need to appeal to two real-valued functions: first, a *probability* function, and secondly, a *value*-function of some kind. To identify a version of the idea that rationality requires one's intentions to maximize expected value which implies Broome's principle as a special case, we must interpret the probability function that is involved in determining the relevant expectation as modelling the rational

agent's *degrees of belief*; and we must interpret the relevant value-function as a measure of how *closely* the fine-grained options approximate to being *what the agent ought to do*.

The idea of modelling a rational agent's degrees of belief by means of probability functions is familiar. The idea of using value-functions to measure how closely such fine-grained options approximate to being what the agent ought to do is less familiar. But it is not too hard to get the hang of this idea. Intuitively it seems clear that we can *compare* the available fine-grained options to each other in terms of how closely they approximate to being what you ought to do. Out of the options that fall short of being what you ought to do, some of these options fall only *slightly* short of being what you ought to do, while others fall *atrociously* far short of being what you ought to do. In other words, of the options that it would be wrong or inappropriate for you to choose, some are more *badly* or *seriously* wrong than others. As I shall say, some are *less choiceworthy* than others.

It seems plausible that there is a way of talking about "reasons for action" on which – at least wherever φ-ing is a fine-grained option – an agent has "most reason" to φ if and only if the agent ought all things considered to φ. So these comparisons of fine-grained options in terms of their degrees of choiceworthiness are effectively equivalent to comparisons of options in terms of *how much reason* there is in their favour.

Moreover, there are reasons for thinking that this notion of choiceworthiness gives us more than just a *ranking* of these options. Specifically, there are reasons for thinking that this notion allows us to make sense of the *cardinal measurement* of choiceworthiness. It seems that we can not only compare *options* in terms of their degrees of choiceworthiness; we can also compare the *differences* in choiceworthiness between options – e.g., we can say that the difference in choiceworthiness between options A_1 and A_2 is a *small* difference, compared to the much *larger* difference between options B_1 and B_2. This supports the view that we can make sense of the cardinal measurement of choiceworthiness.[4] Then we could say that rationali-

[4] A system of four-place relations, comparing the differences between pairs of items with respect to some quantity or value, is known as a *difference structure*. If a difference structure gives a *complete* ranking of all differences with respect to a certain value between *infinitely many* pairs of items, then there is in fact a *unique* interval scale (given an arbitrary choice of a unit and zero point) on which the value in question can be measured. For this point, see Krantz et al. (1971).

ty requires one to have an intention that (out of the relevant set of available alternative intentions) *maximizes expected choiceworthiness*.[5]

To make sense of this notion of expected choiceworthiness, we must suppose that it is rational for the agent to have various *degrees of belief* in various propositions about the *degrees of choiceworthiness* of the relevant options. If these propositions form a partition (that is, it is rational for the agent to be certain that one and no more than one of these propositions is true), the expected choiceworthiness of an option is the weighted sum of its degree of choiceworthiness according to each of these propositions, weighting each degree of choiceworthiness by the degree of belief that it is rational for the agent to have in the relevant proposition.

If the agent has degrees of belief of this sort, then one special case of such degrees of belief is the case in which the agent has the *highest possible* degree of belief in the proposition that option *A* has a *greater* degree of choiceworthiness than *all alternative options* – that is, in effect, the agent is *certain* that *A* is what she *ought* to do. This is the case to which the restricted version of Broome's principle applies.

We may give a more formal representation of this notion of expected choiceworthiness, in the following way. First, let us suppose that there is a set of probability functions, including all and only those probability functions *P* that faithfully represent the degrees of belief that it is rational for the agent to have; let these probability functions be defined over a set of epistemically possible worlds – where these possible worlds are, intuitively, the most specific and detailed propositions that it is rational for the agent to regard as potentially relevant to the decision in question.

Secondly, let us represent these degrees of choiceworthiness by means of a *set* of *real-valued value-functions*. Suppose that every value-function *V* in this set assigns a real number to every relevant world *W* – where *V* assigns a number to each world *W* based purely on the degree of choiceworthiness of the *fine-grained* option that the agent does in the relevant situation in *W*.

Let 'Intend: *A*' stand for the first-personal present-tensed proposition that the agent could express by saying something of the form 'I intend to do *A*'. Then, if *A* and *B* are both fine-grained options, we may say that the intention to do *A* has greater expected choiceworthiness than the intention

---

[5] This proposal is not totally unprecedented. For example, in the context of a discussion of "moral obligation", Peter A. Graham (2010) has suggested (in effect) that a morally conscientious agent will seek to *minimize* her conduct's *expected* degree of moral *wrongness*.

to do B if and only if, for every pair $\langle V, P\rangle$ consisting of one of these value functions and one of these probability functions:

$$\Sigma_W V(W)\ P(W \mid \text{Intend: } A) > \Sigma_W V(W)\ P(W \mid \text{Intend: } B)$$

This notion of expected choiceworthiness can be used to make the following proposal about the intentions that it is rational for an agent to have: When each of the relevant alternative intentions is an intention to do a fine-grained option, rationality requires the agent to have an intention that (out of these alternatives) maximizes expected choiceworthiness in this sense.[6]

This proposal may *look* similar to classical decision theory (according to which rational choices must maximize "expected utility"). In fact, however, if the restricted version of Broome's principle is to be a special case of this proposal, there have to be some crucial differences.

In particular, the restricted version of Broome's principle concerns cases in which the agent is rationally certain that a certain fine-grained course of action is what she ought to do. So, if this version of Broome's principle is to be a special case the proposal that I am making here, then having certain degrees of belief in certain possible worlds must be *equivalent* to having a certain degree of belief in a (normative or evaluative) proposition about which of the relevant options the agent ought to do.

In effect, then, we should think of each world W as an extremely detailed conjunctive proposition, some conjuncts of which are evaluative propositions about the degree of choiceworthiness of the fine-grained act that the agent does in the relevant situation. To get a rough picture of what this amounts to, we might imagine that the relevant evaluative conjunct of this world W is the proposition that the agent might express by saying 'The fine-grained act that I do in this situation is choiceworthy to degree n'. In this case, we could imagine simply that the real number $V(W)$ that the value-function V assigns to W is precisely n.

---

[6] In this definition of an intention's expected choiceworthiness, I have appealed to the *conditional* probability of each world W given the assumption that the agent has the intention in question. For my reasons for defining the notion in this way, see Wedgwood (2011a). For the purposes of this paper, however, this point is not important. These "evidential" conditional probabilities could easily be replaced with a more "causal" notion of probability without affecting my arguments. For my reasons for appealing to the conditional probability of the world given that you *have the intention* (rather than given that you actually *carry out* the intention), see Wedgwood (2011b).

In fact, however, we need to recognize that the precise number that this value-function *V* assigns to each world is really just an arbitrary device for representing the structure of these degrees of choiceworthiness. The relevant possible worlds themselves do not need to assign any real numbers to these degrees of choiceworthiness; they just need to imply enough about these degrees of choiceworthiness so that what the worlds imply about the choiceworthiness of the agent's actions in those worlds can be represented by means of a value-function like *V*.

At all events, we must not think of these possible worlds as encoding only *empirical* uncertainty about non-normative non-evaluative matters of fact; we must think of them as encoding the agent's uncertainty about *normative* and *evaluative* matters as well. The value-function is simply a way of representing a feature of the content of these worlds – specifically, it represents what each world implies about the degree of choiceworthiness of the act that the agent performs in the relevant situation.

In this way, this value function differs crucially from a "utility" function, since the number that a utility function assigns to a world is not determined purely by the content of the world; it is also determined by the agent's subjective preferences, of which this particular utility function is a measure.[7] These preferences can vary independently of the worlds (for example, different agents' utility functions might rank two different worlds in very different ways); so a utility function is clearly not just a way of representing any feature of the content of the worlds.

Admittedly, we have said nothing so far about the precise meaning of the relevant kind of 'ought', or about the nature of this value of "choiceworthiness". So we have not ruled out the suggestion that (like "utility") this value is determined by purely the agent's subjective attitudes. However, even if this value is determined by the agent's subjective attitudes, this proposed principle does not imply that the rational agent's intentions must cohere or harmonize in any way with these subjective attitudes themselves: it requires only that the agent's intentions must cohere with the agent's *de-*

[7] This conception of "expected value" also differs crucially from the conception that appears in Jackson's (1991) "decision-theoretic consequentialism", since the value-function that Jackson appeals to is the value-function that corresponds to the *truth* about morality, whereas in my approach this value-function is simply a way of formulating the *content* of the agent's *degrees of belief* about the relevant options' *degrees of choiceworthiness*.

grees of belief in propositions about the degrees to which the relevant options exemplify this value.

Here is another way of bringing out the distinction between this proposal and classical expected utility theory. Classical expected utility theory is compatible with a strictly *expressivist* and *non-cognitivist* treatment of evaluative and normative language, according to which (at the most fundamental level of analysis) evaluative and normative statements do not express beliefs in ordinary propositions, of the sort that are expressed by ordinary factual statements, but instead express mental states of some fundamentally different "non-cognitive" kind. This sort of expressivist non-cognitivism supports the conclusion that what it is for the mental states that are expressed by these normative and evaluative statements to be rational or justified will ultimately be crucially different from what it is for ordinary factual beliefs to be rational or justified. If this sort of expressivist non-cognitivism is correct, then it is natural to think that we should not model our normative and evaluative attitudes in the same way as our ordinary factual beliefs, as an assignment of degrees of belief across a space of epistemically possible worlds. Instead, we should model these normative and evaluative attitudes as a system of subjective preferences or the like.

By contrast, the proposal that I am making here coheres most straightforwardly with a *cognitivist* and *truth-conditional* interpretation of normative and evaluative statements. According to this sort of cognitivism, even at the most fundamental level of analysis, the meaning of these statements involves a *proposition* – the proposition that gives the truth-conditions of those statements – and in making these statements, speakers express an ordinary attitude of *belief* towards these propositions – an attitude that is of fundamentally the same kind as the attitude of belief that we have towards ordinary factual propositions. This interpretation naturally encourages the view that we have degrees of belief in evaluative and normative propositions, in just the same way as in ordinary factual propositions; and in consequence the relevant epistemically possible worlds, over which our degrees of belief are defined, must be thought of as big conjunctions of both normative and non-normative propositions.

The core of this approach, then, is the idea that the rational agent is guided by her degrees of belief in normative or evaluative propositions about the relevant available options' degrees of choiceworthiness. To that extent, this proposal is in line, not with the Humean tradition, according to which reason is necessarily "the slave of the passions", but rather with

the broadly Aristotelian or Thomistic tradition, according to which the rational will is fundamentally "moved by the intellect".[8]

5.2. Generalizing (ii): Option-individuation

How can the account given above about which *fine-grained* options it is rational to intend be extended into an account of which *coarse-grained* options it is rational to intend?

To capture any requirements that apply to your intentions to do coarse-grained options, we will need a *holistic* constraint on the total set of intentions that you have at the relevant time. Suppose that we can define a notion of the expected choiceworthiness of a whole set of intentions. Then we can say that if you are rational, you will have a whole set of intentions that (out of all relevant alternative sets of intentions) maximizes expected choiceworthiness (in this sense).

Let 'Conj-Intend: P' stand for the proposition that the *conjunction* of all the contents of your intentions is the proposition P. If the conjunction of the contents of one set of intentions S_1 is P and the conjunction of the contents of a second set of intentions S_2 is Q, then S_1 has greater expected choiceworthiness than S_2 if and only if:

$$\Sigma_W V(W)\, P(W \mid \text{Conj-Intend: } P) > \Sigma_W V(W)\, P(W \mid \text{Conj-Intend: } Q)$$

Using this notion of the expected choiceworthiness of a set of intentions, we may now make our most general proposal about what it is rational for an agent to intend: To be rational, an agent must have a *set* of intentions that (out of the relevant alternative sets of intentions) maximizes expected choiceworthiness in this sense.

In other words, the basic idea is this: the agent's intentions must make it rational for the agent to have an expectation for the degree of choiceworthiness that her conduct will exemplify in the relevant situation that is *at least as great* as the expectation that every relevant alternative set of intentions would make it rational for her to have.

It is intuitively clear, it seems to me, that this new principle entails the restricted form of Broome's principle as a special case. That is, this new

---

[8] As Aquinas puts it (cf. *Summa Theologica* IaIIae, 9.1), "*intellectus movet voluntatem*." In the terms that were suggested by Cullity – Gaut (1996), this is fundamentally a "recognitional" rather than a "constructivist" conception of practical reason.

principle guarantees that if you violate the restricted form of Broome's principle, you are irrational.

Suppose that φ-ing is a *super-finely* individuated option, and you are *certain* that in situation S you ought to φ (and Broome's clauses (2) and (3) are met), but at t you do not intend to φ. Then there are two possible cases. In the first case, it is not rational for you to be certain that in this situation you ought to φ. In this case, it is clear that you are being in at least one way irrational.

In the second case, it is rational for you to be certain that in this situation you ought to φ. Given that φ-ing is a super-finely individuated option, the proposition that you ought to φ is by definition equivalent to the proposition that φ-ing is more choiceworthy than all the relevant alternatives. So you are in effect rationally certain that φ-ing is more choiceworthy than all alternatives. In that case, if you are rational, *all* the propositions about the available options' degrees of choiceworthiness in which you have any non-zero degree of belief assign a higher degree of choiceworthiness to φ-ing than to every alternative. At least assuming that Broome's conditions (2) and (3) are met with respect to each of the relevant alternative options, it follows that the intention to φ is the only intention that maximizes expected choiceworthiness. So, in this case, you are rationally required to intend to φ, and if you do not intend to φ you are irrational.

At the same time, this new principle entails the intuitively correct answer to the cases that I put forward in Section 4 as counterexamples to the unrestricted form of Broome's principle. For example, in the case considered in subsection 4.1, no set of intentions that includes the intention to do A will maximize expected choiceworthiness, since on the assumption that you intend to do A, there is too high a risk that your conduct in the relevant situation will involve destroying the world. On the other hand, on the assumption that you intend to do B, there are no such risks of your destroying the world; and so it seems that a set of intentions containing an intention to do B could well maximize expected choiceworthiness, and so could count as a rational set of intentions.

6. Expectations *vs*. beliefs

As we have just seen, in the most general formulation, the principle proposed here says, not that rational agents' intentions are in line with

their *beliefs* about what they *ought* to do, but that their intentions are in line with their *expectations* of *choiceworthiness*.

Beliefs and expectations are crucially different mental phenomena:

a. Each of your beliefs is an attitude towards a *single* proposition.
b. Each of your expectations is determined by your degree of belief in each member of a *partition* of propositions.

There is admittedly one special case in which a belief coincides with an expectation – namely, in the special case in which you are absolutely *certain* of the relevant proposition (in which case the relevant partition of propositions in effect has just one member). Except in this special case, however, beliefs and expectations are importantly different. As David Lewis (1988) taught us, beliefs and expectations behave quite differently in response to new evidence. So long as your initial degree of belief in a proposition $p < 1$, new evidence can lower your degree of belief in p below any threshold t; that is, in effect, new evidence can *deprive* you of having any belief in p. By contrast, if you are rational, new evidence will never deprive you of having any expectation of choiceworthiness for the intention to do A.

So, according to the proposal that I am making here, the fundamental account of rationality is that the rational agent's intentions are in line with her expectations of choiceworthiness, not with her beliefs about what she ought to do.[9]

Some philosophers will be inclined to object that I am underestimating the importance of beliefs about what one ought to do. In particular, some of these philosophers will object along the following lines. According to the principle of Section 5, whenever there is a unique set of intentions that maximizes expected choiceworthiness, you are rationally required to have those intentions. Arguably, the notion of a "rational requirement" is a kind of 'ought' – specifically, it is a "subjective 'ought'", in the sense that what the agent "subjectively *ought*" to do is determined by facts about the evidence or information that is available to that agent (not by facts about the external world of which the agent is ignorant). If that is right, then in all of these cases, there is a sense of 'ought' such that the agent *ought* in that sense to have the intentions that she is rationally required to have. So, in

[9] One philosopher who appreciated this point more than twenty years ago, ironically enough, was John Broome (1991), in his commentary on Lewis (1988).

every one of these cases, there is a kind of 'ought' – the 'ought' of rational requirement – that the rational agent's intentions will conform to.

Moreover, some philosophers might think that because this kind of 'ought' – the 'ought' of rational requirement – is fixed by the degrees of beliefs that it is rational for the agent to have, there is no room for any real difference here between the claim that the agent "ought" in this sense to intend to φ (that is, the agent is rationally required to intend to φ) and the claim that the agent *rationally believes* that she "ought" in this sense to intend to φ. So these philosophers would think that a principle very similar to the unrestricted form of Broome's principle is true – namely, the principle that to be rational, one must intend to φ whenever one rationally believes that one ought in this sense to intend to φ.

As tempting as this line of thought may seem to some philosophers, it cannot be reconciled with the Williamson-inspired point that no non-trivial 'ought' is rationally luminous. Since the notion of a "rational requirement" is a kind of 'ought', rational requirements are not luminous either: cases can always arise where it is *true* that you are rationally required to intend to φ, but it is impossible for you rationally to believe that you are rationally required to intend to φ. So, it seems that for *every* sense of 'ought', there is a gap between the proposition that you ought, in this sense, to intend to φ, and the proposition that you *rationally believe* that you ought, in this sense, to intend to φ. Cases can arise where the first proposition is true and the second proposition is false.

As a result, the principle that says that for you to be rational, your intentions must be in line with your rational beliefs about what intentions you are rationally required to have is at best significantly *narrower* – that is, covers a significantly smaller range of cases – than the principle that I am advocating here, according to which for you to be rational, your intentions must actually maximize expected choiceworthiness. If the more general principle is correct, as I am advocating here, then it seems that the narrower principle has no interest except as a special case of that more general principle.

Moreover, many philosophers would accept that it is possible for you to have a rational but *false* belief in the proposition that you are rationally required to intend to φ. (For example, if a suitable oracle pronounces that you are rationally required to intend to φ, perhaps you could be rationally required to have a high degree of confidence in the proposition that you are rationally required to intend to φ, even if on this one occasion, the oracle's

pronouncement is actually false.) If this is indeed possible, then these two principles are not just different but *inconsistent* with each other. In that case, since I am advocating the principle that to be rational, your intentions must maximize expected choiceworthiness, I would be committed to denying the principle that to be rational, your intentions must be in line with your rational beliefs about what intentions you are rationally required to have.

In general, then, it is not rational agents' *beliefs* about what is rationally required of them that fundamentally guide their deliberations. It is these *rational requirements themselves* – and the facts about the agents' degrees of belief in propositions about the available options' degrees of choiceworthiness, on which these rational requirements supervene – that will guide the agents' deliberations.

This point helps us to understand the precise sense of 'ought' that appears in the true instances of Broome's principle. This sense of 'ought' is closely connected to the notion of "choiceworthiness": as I put it above, choiceworthiness is precisely a measure of how closely each of the available options approximates to being what the agent ought, all things considered, to do in the relevant situation. Thus, the very concept of choiceworthiness guarantees that for any fine-grained option A, the agent *ought* in the relevant sense to do A if and only if A is the *most choiceworthy* option available to that agent in the relevant situation.

According to the principle that was proposed in Section 5, uncertainty is handled by means of the *degrees of belief* (and the sets of probability functions that represent those degrees of belief) that are involved in defining the notion of *expected choiceworthiness*. For this reason, it would be *double-counting* if the concept of 'choiceworthiness' that appears in this proposed principle also took account of uncertainty. So, this concept of "choiceworthiness" must be a *maximally objective* normative notion: the truth about the relevant available options' degrees of choiceworthiness depends on the objective facts of the agent's situation, which may include facts that the agent is not even in a position to know at the relevant time.

I argued at the end of the previous section that the restricted version of Broome's principle is just a special case of the principle that was proposed in that section. Given the fundamental connection between the concepts expressed by these uses of 'ought' and 'choiceworthiness', it follows that the sort of 'ought' that appears in this restricted version of Broome's principle must also be a maximally objective 'ought' – the kind of 'ought' for which

what one ought to do may depend on facts that one is not in a position to know at the relevant time.[10]

7. The fundamental principle of rationality

Someone might object to the principle that was proposed in Section 5, along the following lines. It seems possible for there to be an agent who is capable of rational choices and intentions, but has no degrees of belief in *any* propositions about the relevant options' degrees of choiceworthiness. How could such an agent have a set of intentions that maximizes expected choiceworthiness? And if it is not possible for an agent to have a set of intentions that maximizes expected choiceworthiness, how could the agent be rationally required to have such intentions?

My definition of expected choiceworthiness appeals to a set of probability functions – specifically, the set of probability functions that models the degrees of belief that it is rational for the agent to have. This objection to the principle that I proposed in Section 5 fails, because there can *be* degrees of belief that it is rational for the agent to have, even if the agent does not *actually* have these degrees of belief.

According to most theories of rationality, the degrees of belief that it is rational for an agent to have at a given time are determined by such things as the evidence that the agent has at that time, or by the mental states and mental events that are present in the agent's mind at that time. Even if the agent does not actually have these degrees of belief, it is presumably still possible for the agent to be guided, in forming and revising her intentions, by this body of evidence, or by these mental states and mental events. In this way, it could be that it is no accident that the agent has a set of intentions that maximizes expected choiceworthiness – if the agent is guided, in forming and revising these intentions, by whatever evidence or mental states determine the degrees of belief that it is rational for the agent to

[10] At the same time, it is easy to explain why the *assertions* that we make about what agents ought to do will often involve a less objective kind of 'ought'. When we make assertions, we are generally highly confident of the truth of what we say. We are often not very confident of the propositions involving this objective 'ought'. Still, even if this objective 'ought' is often not used in our assertions, it may nonetheless appear in propositions towards which we have partial degrees of belief.

have in the relevant propositions about the options' degrees of choiceworthiness.

Once the principle proposed in Section 5 is clarified in this way, it becomes clear that if this principle is true, it is not just one principle of practical rationality among many. It is in a sense the *fundamental* principle of rationality. This is because this principle applies quite *generally*, to all cases of rational intention; so this principle conflicts with all other proposed principles of rational intention, except for those that are implied by it.

For example, at least on most interpretations of what "utility" is, the principle proposed in Section 5 conflicts with the principle that rational intentions must maximize expected utility. The reason for this is that on many common interpretations of "preference", it is possible for a rational agent to have a "preference" for *A* over *B*, even if the expected choiceworthiness of *B* is greater than that of *A*. The dominant interpretation of "utility" – ever since John von Neumann and Oskar Morgenstern (1944) – is simply as a measure of the relevant agent's "preferences". So (on those common interpretations of "preference") it is possible for *A* to have greater expected choiceworthiness than *B* even if *B* has greater expected utility than *A*. It follows that the principle proposed in Section 5 is inconsistent with the view that rational intentions must maximize expected utility.

There are still some ways in which this principle needs to be clarified. In particular, we need to figure out exactly how to understand the "relevantly available alternative sets of intentions" that are mentioned in the principle. What is it for a set of intentions to be one of these "relevantly available alternative sets"? This point clearly needs to be clarified if we are to understand the exact implications of this principle.[11]

Subject to these clarifications, however, our search for an anti-*akrasia* principle that can handle cases of uncertainty and intentions involving coarsely individuated options seems to have led us to the fundamental principle of practical rationality.

[11] I have tried to explore this question in a little more detail elsewhere; see Wedgwood (2011b).

References

BRATMAN, M. (1988): *Intentions, Plans, and Practical Reasoning*. Cambridge, MA: Harvard University Press.

BROOME, J. (1991): Desire, Belief and Expectation. *Mind* 100, 265-267.

BROOME, J. (2013): *Rationality through Reasoning*. Oxford: Wiley-Blackwell.

CULLITY, G. – GAUT, B. (eds.) (1996): *Ethics and Practical Reason*. Oxford: Clarendon Press.

GRAHAM, P. (2010): In Defense of Objectivism about Moral Obligation. *Ethics* 121, 88-115.

JACKSON, F. (1991): Decision-Theoretic Consequentialism and the Nearest and Dearest Objection. *Ethics* 101, 461-482.

KOLODNY, N. (2005): Why Be Rational? *Mind* 114, 509-563.

KRANTZ, D. H. – LUCE, R. D. – SUPPES, P. – TVERSKY, A. (1971): *Foundations of Measurement I: Additive and Polynomial Representations*. London: Academic Press.

LEWIS, D. K. (1988): Desire as Belief. *Mind* 97, 323-332.

MORGENSTERN, O. – VON NEUMANN, J. (1944): *Theory of Games and Economic Behavior*. Princeton, NJ: Princeton University Press.

PARFIT, D. (2011): *On What Matters*. Vol. 1. Oxford: Oxford University Press.

REGAN, D. (1980): *Utilitarianism and Cooperation*. Oxford: Clarendon Press.

WEDGWOOD, R. (2011a): Gandalf's Solution to the Newcomb Problem. *Synthese*, Online First: 15 March, 1–33. DOI: 10.1007/s11229-011-9900-1.

WEDGWOOD, R. (2011b): Instrumental Rationality. *Oxford Studies in Metaethics* 6, 280-309.

WILLIAMSON, T. (2000): *Knowledge and its Limits*. Oxford: Clarendon Press.