Epistemic Teleology: Synchronic and Diachronic

Ralph Wedgwood

According to a widely-held view of the matter, whenever we assess beliefs as “rational” or “justified”, we are making *normative* judgments about those beliefs. In this discussion, I shall simply assume, for the sake of argument, that this view is correct. My goal here is to explore a particular approach to understanding the basic principles that explain which of these normative judgments are true. Specifically, this approach is based on the assumption that all such normative principles are grounded in facts about *values*, and the normative principles that apply to beliefs in particular are grounded in facts about *alethic* value – a kind of value that is exemplified by believing what is true and not believing what is false. In this essay, I shall explain what I regard as the best way of interpreting this approach. In doing so, I shall also show how this interpretation can solve some problems that have recently been raised for approaches of this kind by Selim Berker, Jennifer Carr, Michael Caie, and Hilary Greaves.

1. The value-based conception of normativity

In this discussion I shall explore what I call the “value-based conception of normativity”. According to this conception, *orderings* or *rankings* of *states of affairs* – rankings of the sort that could be expressed in English by means of comparative evaluative terms like ‘better’ and ‘worse’ – lie at the heart of normativity.

According to this conception, all of the key terms used to assess beliefs – such as ‘rational’ and ‘justified’ and the like – can be used express *evaluative* concepts. When the terms are used in this way, for a belief to be “rational” or “justified” is for it to be a belief that is *good* in a certain way, while for a belief to be “irrational” or “unjustified” is for it to be a belief that is in a certain corresponding way *bad*.[1] (In this discussion, I shall assume that, even if the terms ‘rational’ and ‘justified’ differ in their connotations or pragmatic features, in the contexts that are pertinent to our purposes they express the same concept.)

Like every other kind of goodness, rationality and justifiedness come in *degrees*: one belief can be more irrational, or more unjustified, than another; and for one belief to be “more irrational” than a second belief is for the first belief to be in a certain way *worse* than the second. This relation of *being more irrational than*, like the relation of being *worse than*, is irreflexive, asymmetric, and transitive. In this way, the relation of *being at least as irrational as*, just like the relation of being *at least as bad as*, provides a *partial ordering* of the items that it relates.

Our focus here is on the use of terms like ‘rational’ and ‘justified’ to assess beliefs. More precisely, when we say that “a belief” is rational, we are really predicating rationality of the relevant thinker’s *having* the relevant belief at the time in question – that is, in effect of the *state of affairs* that consists in the thinker’s having that belief at that time. In this way, interpreting ‘rational’ as expressing an evaluative concept is consistent with the assumption that the basic function of evaluative concepts is to evaluate and rank alternative states of affairs. To simplify our discussion, I shall just speak of the

[1] It is important here to realize quite how many ways of being good – that is, how many different values – there are. Most philosophers working in ethical theory have dramatically underestimated both how many such ways of being good there are, and the extent to which terms like ‘good’ are context-sensitive. Even Thomson (2009), who rightly emphasizes the many ways in which the term ‘good’ is used, in the end gives what seems to me an unduly limited catalogue of all the different concepts that the word can express.

rationality of beliefs and the like. But what really counts as "rational" or as "irrational" is the thinker's having the belief at the time in question.

In addition to being used to assess beliefs, terms like 'rational' and 'justified' can also be used to assess other items besides beliefs:

a. They can be used to assess whole *systems* of mental states, such as belief-systems, or collections of plans and intentions.

b. They can be used not only for mental states, but also for *mental events* – and especially for mental events in which the thinker *forms* a new mental state of some kind, or *revises* (or *reaffirms*) some of her old mental states. These mental events include *judgments* – events in which we form a new belief or reaffirm an old belief – and also *choices* or *decisions* – events in which we form, revise or reaffirm our plans and intentions.

This distinction, between assessing the rationality of static mental states like beliefs and intentions, on the one hand, and assessing the rationality of mental events in which we change or revise our mental states, on the other, is one way of understanding the distinction between *synchronic* and *diachronic* rationality. As we shall see later on, however, there is another deeper and more important way of understanding this distinction as well.

Although I shall assume that the evaluative concepts expressed by terms like 'rational' and 'justified' rank state of affairs of all these kinds, I shall not assume that the states of affairs ranked by these evaluative concepts include states of affairs that consist in the thinker's simply *lacking* any doxastic attitude (even suspension of judgment) towards a certain proposition. Perhaps such states of affairs – consisting, not in doxastic attitudes, but in the *absence* of such attitudes – are not evaluated by these concepts at all.[2] I shall also assume that this limitation on the domain of states of affairs that are evaluated by these concepts is non-trivial, since thinkers may be totally *attitudeless* towards certain propositions (even if these propositions are logically complex propositions built up out of atomic propositions towards which the thinkers *do* have attitudes).

Besides this way of using terms like 'rational' and 'justified' as evaluative terms that apply to mental states and mental events, there is also another way of using these terms. When using the terms in this other way, we speak of the "*requirements* of rationality" and of what is "rationally required" of a thinker at a particular time; and we can also speak of what is *rationally permissible* – that is, consistent with all rational requirements – for a thinker at a time.

The connection between these two ways of using terms like 'rational' or 'justified' is, I propose, quite simple. The terms 'required' and 'requirement' mean just what they sound like: they mean what is *needed* or *necessary*. To say that something is necessary, as the modal logicians taught us many years ago, is equivalent to universally quantifying over possibilities: something is necessary if and only if it is the case in *all* the possibilities in the relevant domain.

As virtually all philosophers of language and semanticists agree, the domain of possibilities that we are quantifying over is determined by the *context* in which terms like 'necessary' and 'required' are used. I shall follow tradition in representing these possibilities by talking about *possible worlds*. So, whenever we speak of what is "necessary" or "required", something in the context of our conversation determines a domain of possible worlds. To simplify matters, let us assume that this is simply the domain of worlds that we are *thinking of* in the context. For example, we may be thinking of the

[2] In fact, towards the end of this section, I shall argue, against Carr (2015), that there may be no way of comparing the degree of irrationality of one belief-system that is defined over one set of propositions with the degree of irrationality of a second belief-system that is defined over a second *distinct* set of propositions. See note 6 below.

possible worlds that are in some way *available* to you at the present time, through the way in which you avail yourself of the opportunities that you have for exercising your cognitive capacities at this time. Moreover, in each context where we speak of the "requirements of rationality", there will be some items – call them the "relevant items" – that we are interested in assessing for rationality in that context. For example, we may often be interested in assessing the possible total belief-systems that you might have right now. Then the relevant item in each world will be the total belief-system that you have now in that world.

So, suppose that the context has determined a domain of worlds, and a relevant item that is up for assessment in each of these worlds. Then it may be that within this domain of worlds, there are some worlds where the relevant item is *maximally rational* – that is, *no less rational* than the corresponding relevant item in any other world. Just to have a label, let us call these the "rationally favoured" worlds. Then, in general, something is rationally required if and only if it is the case at all the rationally favoured worlds. For example, suppose that we are thinking of the possible worlds that are available to you now, and interested in assessing the possible total belief-systems that you might have now. Then what is rationally required of you now is everything that is the case at all the worlds that both (a) are available to you now, and (b) are worlds where the total belief-system that you have now is no less rational than the total belief-system that you have at any other available world.

So long as there always are worlds within the relevant domain where the relevant item is maximally rational, this account guarantees that the operator 'It is rationally required that …' obeys all the principles that apply to deontic modals like 'ought' or 'should' according to standard deontic logic – that is, the modal system KD. This makes it plausible, it seems to me, that the notion of a "rational requirement" can be expressed by speaking, in a certain way, of how we "should" or "ought to" think.[3]

However, in addition to assessing beliefs as "rational" or "justified", we also assess beliefs as "right" or "wrong", "correct" or "incorrect"; and these terms typically seem to express different concepts from those that are most commonly expressed by 'rational' or 'justified'.

It is not standard in ordinary English to speak of one belief's being "less correct" than another. But it still seems possible to make sense of the idea of degrees of incorrectness or wrongness. At least, we can qualify terms like 'wrong' and 'incorrect' by a range of modifiers: one thing might be "utterly wrong" (or "deeply wrong" or "egregiously wrong"), while another thing might only be "slightly wrong" (or "not [too] far wrong"). So these terms seem capable of functioning as evaluative terms: a correct belief is in a way a *good* belief, while an incorrect belief is in a corresponding way a *bad* belief; and it seems that this value of correctness and incorrectness, like all other values, comes in degrees.

There may be many different concepts that terms like 'correct' can express. But in this discussion, I shall explore the idea that these terms can express a kind of *alethic value* that beliefs can possess. Consider a belief – that is, a state of affairs consisting in a certain thinker's believing a certain proposition *A* at a certain time. If 'correct' expresses this kind of alethic value, then this belief is correct just in case *A* – the proposition that is believed – is *true*; and it is incorrect just in case *A* is not true. Moreover, with this kind of alethic value, we can understand the idea of degrees of incorrectness in the following way: if *A* is *true*, then the greater the degree of confidence with which the thinker

[3] Compare the semantics of deontic modals that was originally developed by the deontic logicians of the 1960s and 1970s – such as David Lewis (1973), usefully summarized by Lennart Åqvist (1984) – and later applied to the empirical data of language use by linguists such as Angelika Kratzer (2012).

believes *A*, the *less* incorrect the belief is; if *A* is *not* true, then the greater the degree of confidence with which the thinker believes *A*, the *more* incorrect the belief is.[4]

This, then, is one thing that it could mean to say that one belief-state towards a proposition *A* is "more incorrect" than another such belief-state. When the term is used in this way, it also seems to make sense to say that while one belief-state b_1 is *slightly* more incorrect than another such belief-state b_2, the belief-state b_3 is *much* more incorrect than another belief-state b_4. That is, we can compare the *differences* in degrees of incorrectness between *pairs* of such belief-states.[5] This suggests that it may in principle be possible to *measure* a belief-state's degree of incorrectness, by means of an appropriate *scoring rule*. We shall return to the question of what this scoring rule might be like in Section 5 below.

In addition to assessing individual beliefs as correct or incorrect, we can also assess whole belief-systems. If there is indeed a scoring rule that measures the degree of incorrectness of each individual belief-state, then it is tempting to think that the way to assess whole belief-systems is by some kind of *weighted sum* of the degrees of incorrectness of the individual belief-states that make up this belief-system.

The reason for thinking that the way to assess a whole belief-system is by means of a *weighted* sum of the individual belief-states' degrees of incorrectness is just that some of the propositions that are the objects of these individual belief-states may be *more important* than others. We should allow that these weights are determined holistically by the whole set of propositions in which the thinker has belief-like or doxastic attitudes of any kind. In that case, these weights would only be defined for each set of propositions; and there may not be any way of comparing the incorrectness of one belief-system, which consists of attitudes towards *one* set of propositions, with another belief-system, which consists of attitudes towards *another* set of propositions. However, we may not need to make such comparisons anyway: it may be enough if the relevant kind of alethic value yields only a *partial* ordering of the items to which it applies.[6]

This, then, is the approach that I shall explore in this discussion. This approach is based on the assumption that all normative terms that we use to assess beliefs express either (a) evaluative concepts – which stand for values that come in degrees, and yield a partial ranking of the items to which they apply – or else (b) concepts of the "requirements" of such values, of the kind that I have explained, which can be expressed by deontic modals like 'should' and 'ought'. Two sets of such evaluative concepts stand out as particularly central – namely, the concepts that can be expressed by

[4] For the seminal discussion of this way of evaluating beliefs, see Joyce (1998).

[5] In other words, these degrees of incorrectness form a *difference structure*. For an explanation of why this supports the idea of cardinal measurement, see Krantz et al. (1971: 150-2 and 157-8).

[6] Whether one should "expand" one's belief-system by including a doxastic attitude towards a proposition that one previously had no attitudes towards whatsoever is not in my view a question about *epistemic* rationality at all. Typically, exogenous mental events that are outside one's control simply *compel* one to start having attitudes towards a proposition. (For example, one might have a *sensory experience* that compels one to start having an attitude towards a proposition that forms part of the content of the experience; or a new proposition might simply *occur* to one, as a hypothesis, which in effect would also compel one to start having some attitude towards the proposition.) Even suspending judgment about a proposition, or wondering whether it is true, is a kind of broadly doxastic attitude towards the proposition; and such an attitude may well be less rational than having a more definite level of confidence in the proposition. For this reason, then, I disagree with the arguments of Carr (2015) that an account of epistemic rationality needs to include a story about when such "epistemic expansions" are rational.

terms like ‘rational’ or ‘justified’, and the concepts that can be expressed by terms like ‘right’ or ‘correct’. We shall explore the approach that is based on the assumption that one fundamental kind of assessment of beliefs, which can be expressed by terms like ‘right’ and ‘correct’, is in terms of this kind of *alethic* value.

2. ‘Consequentialism’, ‘utility’, ‘decision theory’, ‘teleology’

Different philosophers have deployed different terminology to refer to this value-based approach to epistemology. Unfortunately, much of this terminology is at least potentially misleading.

Thus, some philosophers, such as Selim Berker (2013), speak of “epistemic consequentialism”. But consequentialism is best understood as the doctrine that the value or normative status of the relevant states of affairs (whether these states of affairs be *acts*, or *beliefs*, or anything else) is derivative from the value of these states of affairs’ *total consequences*.

There are various ways of thinking of “the total consequences” of a state of affairs. One way, for example, is in terms of *counterfactual* or *subjunctive* conditionals: the total consequence of a state of affairs consists of the conjunction of everything that would be the case if the state of affairs in question obtained. Alternatively, one might think that the total consequence of a state of affairs is nothing short of a whole possible world. Then we might propose that the way in which the value of states of affairs is derivative from the value of possible worlds is to be explained in terms of *conditional objective chances*: the value of a state of affairs is the weighted sum of the values of all possible worlds that are compatible with that state of affairs, weighting the value of each world by the conditional objective chance of the world, conditional on that state of affairs.

At all events, the value-based approach that I have sketched does not assume that the value of beliefs, or belief-systems, or events of belief-revision, is in any way derivative from the value of their total consequences. On the contrary, each of these items can instantiate the value of correctness, or the value of rationality, *in its own right*: its value is not derivative from the value of “consequences” in any interesting sense. To put what is essentially the same point in other words, according to this approach what are relevant are not the values that these items (such as beliefs or belief-systems or belief-revisions) *promote*, but the values that these items themselves *instantiate*.[7]

Some other philosophers, such as Richard Pettigrew (2013) refer to this value-based approach as “epistemic utility theory”. But this too is misleading, because “utility” strictly speaking is a measure of *subjective preference*.[8] Fundamentally, the value-based approach to normativity is quite general: it could invoke values of any kind – not just this one very specific kind of value, “utility”. Specifically, I am focusing on the two crucial values that I have called *rationality* and *correctness*. It is doubtful whether it would be plausible to identify either of these values with utility. At all events, this is certainly not something that we should assume at the outset. So it is important not to be misled by the terminology of “epistemic utility”.

Other philosophers, such as Hilary Greaves (2013), have spoken of “epistemic decision theory”. This description is also potentially misleading. Although the term can be used more broadly, in the strict sense, decision theory, or rational choice theory, as its name suggests, is concerned with *decisions* or *choices*. As I use the term, a choice or decision is a mental event in which an agent forms, revises, or reaffirms her *plans* or *intentions* about how to act. In this sense, a belief is not a decision – nor is there

[7] For a discussion of the importance of this distinction between what *promotes* values and on what *instantiates* these values, see Pettit (1991, 231).

[8] For this point, see especially Broome (1991).

any reason to think that our beliefs are typically, or indeed ever, chosen by us, in the way in which our intentional actions are chosen. Decision theory is concerned with rational choices or decisions; we are concerned here with a different topic – namely, rational belief and rational belief revision. So the term 'decision theory' also seems best avoided here.

A slightly better term to use in this context is "teleology". Even this term is potentially misleading too, because it suggests an *aim* or *telos* – and many philosophers seem to think that every aim (or *telos*) must be either intended by an agent or else an evolutionary proper function of some kind. I shall not argue here that beliefs have an "aim" of either of these kinds. We can, however, interpret this talk of an "aim" more metaphorically, so that it refers only to a fundamental value. This is how John Rawls (1971, 21) understood 'teleology', as referring to the view that "the good is prior to the right". Taken in this sense, the terminology of "epistemic teleology" seems acceptable.

3. Trade-offs

I have proposed that the degree of incorrectness of a whole belief-system is some kind of weighted sum of the degrees of incorrectness of the individual belief-states that make up this belief-system. This proposal about the relation between the value of individual beliefs and the value of whole belief-systems obviously allows for a kind of trade-off. One belief-system b_1 can be, overall, better than another such system b_2, even though there is some particular proposition A such that the attitude towards A involved in b_1 is worse than the attitude towards A involved in b_2.

Selim Berker (2013) has objected to such trade-offs, building on some objections to reliabilism that were originally made by Roderick Firth (1981). In objecting to these trade-offs, Berker correctly notes that the idea that such trade-offs are possible is structurally analogous to a famous feature of classical utilitarianism. According to utilitarianism, the goodness of a possible world is measured by the total sum of welfare that it contains. So a world in which there are ten individuals, each of whom has a welfare level of 90, is less good than a world in which there are just these ten individuals, and nine of them have a welfare level of 100, and one of them has a welfare level of 10. This implication of utilitarianism is troubling enough; but given utilitarianism's commitment to act-consequentialism, it also follows that an *act* whose consequence is the first world is less good than an act whose consequence is the second world, and if no other acts are available, the agent *ought* to perform the second act rather than the first.

Philosophers who (like me) reject utilitarianism offer different diagnoses of the fundamental flaw in utilitarianism that is revealed by this troubling implication. According to one diagnosis, the fundamental flaw lies in utilitarianism's conception of how the goodness of a whole world is related to the welfare-levels of the individuals that it contains – a conception that wrongly attaches no importance to *equality* between different individuals' welfare levels. A second much more radical diagnosis is that ethical theory should not countenance such trade-offs between the interests of distinct individuals at all.

The second diagnosis seems, at least in my judgment, far too extreme. There are plenty of cases in which we must countenance such trade-offs. A world w_1 where there is one fabulously well-off individual x_1 and nine wretchedly badly-off individuals x_2 …. x_{10} is clearly inferior to a world w_2 where x_1 is slightly less well off than in w_1 and x_2 …. x_{10} are dramatically better off than in w_1. But this comparison between the worlds w_1 and w_2 involves a trade-off between the loss to the fabulously well-off individual x_1 and the gains to the wretchedly badly-off individuals x_2 …. x_{10}. Whatever insight there may be in Rawls's (1971, 27) famous remark "Utilitarianism does not take seriously the distinction between persons", it should not be interpreted as equivalent to this extreme second diagnosis, which crazily rejects all trade-offs whatsoever.

So, it seems to be the first diagnosis of this fundamental flaw in classical utilitarianism, and not the second diagnosis, that is correct here. In general, whenever a value is exemplified both by a complex state of affairs, and also by some of the simpler states of affairs that are the constituents of that complex state of affairs, some trade-offs seem unavoidable. The only question is: Which trade-offs? The thesis that there are no trade-offs at all is simply incredible.

In Section 1, I proposed that the degree of incorrectness of a whole belief-system is simply a weighted sum of the degrees of incorrectness of the individual belief-states that constitute that system. I also proposed that these weights are determined by the set of propositions that are the objects of the belief-states in question – and that in consequence there is no way of comparing the degree of incorrectness of a belief-system that is defined over one set of propositions with the degree of incorrectness of a belief-system that is defined over a different set of propositions.

It follows from these proposals that one belief-system b_1 can, overall, have a better degree of incorrectness than another belief-system b_2, even though there is some particular proposition A_1 such that the attitude towards A_1 involved in b_1 has a worse degree of incorrectness than the attitude towards A_1 involved in b_2. For example, perhaps b_1 involves an incorrect attitude towards A_1, but correct attitudes towards a large number of other propositions $A_2, \ldots A_n$, while b_2 involves a correct attitude towards A_1, and incorrect attitudes towards $A_2, \ldots A_n$. (This could be the case, for instance, because A_1 is false and $A_2, \ldots A_n$ are true, and b_1 involves total disbelief in all these propositions – including the false proposition A_1, and also all the true propositions $A_2, \ldots A_n$ – while b_2 involves confident belief in all these propositions.)

In this case, however, surely b_1 *is* better, in terms of its overall degree of incorrectness, than b_2. This trade-off seems positively plausible, at least to me. Still, someone who – like Berker – wishes to press this objection might insist that it is implausible to say that the thinker *should* have b_1 rather than b_2: surely, the thinker should not believe a false proposition like A_1 just to gain true beliefs in $A_2, \ldots A_n$?

I concede that it does sound strange to say that the thinker should *revise* her beliefs by shifting from b_2 to b_1. But I suggest that there are two reasons for why it sounds strange, which are both quite compatible with the proposals that I have made above. First, in any normal case, when a thinker is in the belief-system b_2 and is considering the propositions $A_1, \ldots A_n$, there will be other options available besides just sticking with b_2 or shifting to b_1. For example, the option b_3 of *disbelieving* the false proposition A_1, and confidently believing all the true propositions $A_2, \ldots A_n$, is also available, and that option is clearly preferable to both b_1 and b_2. If an option like b_3 is available, then plainly the thinker should not shift to b_1.

Secondly, the question about whether we should *revise* our beliefs in a certain way is in principle different from the question of whether we should *hold* a certain system of beliefs. In fact, it seems to me that – when it comes to degrees of incorrectness (as opposed to degrees of irrationality) – the correctness of a belief-revision is simply identical to the degree of incorrectness of the belief-system that results from the belief-revision. However, it seems hard to hear the question of whether we "*should*" revise our beliefs in a certain way as a question about the requirements of *correctness* – as opposed to as a question about the requirements of *rationality*. As I shall explain later on, just because b_1 has a better degree of incorrectness than b_2, it certainly does not follow that the transition from b_2 to b_1 is rational; indeed, this transition may well be grossly irrational.

In these ways, then, I can explain why it sounds bizarre to say that the agent should shift from b_2 to b_1 in this case. Since it is hard to see any other reason for doubting that b_1 has a better degree of incorrectness than b_2, I conclude that the proposals that I have made in Section 1 are immune to the objections that have been raised by Berker.

4. The probabilistic connection between correctness and rationality

So far, I have discussed these two key evaluative concepts – rationality and correctness – quite separately, without exploring any of the relations between them. I have also not said anything about how to measure the *degree of irrationality* that is exemplified by a belief (or system of beliefs) or by a mental event in which we revise (or form or reaffirm) our beliefs. One of the most promising ideas that has been explored in recent epistemology – especially in a series of seminal works by James M. Joyce (1998 and 2009) – focuses on the possibility of a fundamental connection between rationality and the kind of alethic value that I am referring to with the term 'correctness'.

Suppose that there is some crucial notion of what is *epistemically possible* for a thinker at a given time – a kind of possibility that at least ideally, rationally should be guiding the thinker at that time. There are several features that this notion of epistemic possibility may be presumed to have. For example, if *A* is a logical truth, then no proposition incompatible with *A* is epistemically possible – in that sense, all logical truths are epistemically necessary. It may also be that all conceptual or analytic truths are epistemically necessary in the same way (even if, like 'If *B*, then Actually *B*', they are not metaphysically necessary).

This kind of epistemic possibility can be used to define some useful further notions. Consider two belief-systems, b_1 and b_2. Suppose that it is epistemically *impossible* for b_1 to have a better degree of incorrectness than b_2, but epistemically *possible* for b_2 to have a better degree of incorrectness than b_1. Then we can say that b_2 at least *weakly dominates* b_1. If in addition it is epistemically *necessary* that b_2 must have a better degree of incorrectness than b_1, we can say that b_2 *strongly dominates* b_1.

Suppose that a belief-system b_1 is strongly dominated by some available alternative belief-system b_2; and suppose that b_2 is not even weakly dominated by any other available belief-system in this way. In this case, it seems that b_1 cannot be maximally rational. There will be some other belief-system distinct from b_1 – perhaps b_2, or perhaps some third available belief-system b_3 – that is maximally rational.

As Joyce (1998) has shown, this dominance principle has several important implications. Given certain assumptions about how these degrees of incorrectness are to be measured, and about the relevant notion of epistemic possibility, this principle entails that, so long as enough belief-systems are available, no probabilistically incoherent belief-systems can be maximally rational. In this way, the dominance principle may explain why rationality requires having a probabilistically coherent belief-system.

I shall not dispute the dominance principle here. It is, however, a very *weak* principle: it only applies in a very narrow range of cases – namely, in cases in which a belief-system is strongly dominated by some available alternative. There seem to be forms of irrationality that cannot be captured in this way. (For example, the belief-systems that make it impossible to learn anything from experience, or from induction, seem to be irrational; but it is doubtful whether they are strongly dominated by any available alternative.) We need to find a more general principle – presumably, a principle that entails the dominance principle as a special case.

A more general principle will not be hard to find if we may suppose that the relevant kind of epistemic possibility comes in *degrees*: that is, some propositions are *more possible* than others. In picturesque terms, we might suppose that the epistemically possible worlds form a *space*, and some sets of possible worlds take up larger proportions of the space than others. These proportions between subsets – or in more picturesque terms, sub-regions – of the whole space of epistemically possible worlds could then be measured by means of a *probability function*.

If the space of epistemically possible worlds does indeed have this structure, then the idea implicit in the dominance principle – that for every rational thinker and every time, there is a kind of epistemic possibility that rationally should be guiding the thinker at that time – can be developed into the more general idea that for every rational thinker and every time, there is *probability function* that should be guiding the thinker at that time. Just to give it a label, let us call this probability function the *rational* probability function for this thinker at this time.

Several further refinements could be introduced into this proposal. For example, perhaps there is not always a *unique* probability function that rationally should guide each thinker at each time; perhaps, sometimes, it is only a big *set* of such probability functions that rationally should be guiding the thinker at the time. However, I shall ignore these complications here: to fix ideas, I shall assume that there is always a unique rational probability function for each thinker and time.

Suppose that it is indeed the case that for every thinker and time, there is a rational probability function of this kind; and suppose also, as I suggested in Section 1, that the degrees of incorrectness discussed above can be *measured* on an interval scale, by means of an appropriate scoring rule.[9] Then we can make sense of a belief-system's *expected* degree of incorrectness – where a belief-system's expected degree of incorrectness is the weighted sum of its epistemically possible degrees of correctness, weighting each of these degrees of incorrectness by the appropriate "rational probability" of the belief-system's having that degree of incorrectness.

My proposal is that each belief-system's degree of irrationality is determined by how its expected degree of incorrectness compares with that of the available alternatives. According to this rational probability function, there will at least one belief-system that optimizes – that is, minimizes – expected incorrectness (that is, no available alternatives to this belief-system have a lower expected degree of incorrectness); this belief-system's expected degree of incorrectness is the optimal available expected degree of incorrectness. The *greater* the extent to which, according to this rational probability function, a belief-system's expected degree of incorrectness *falls short* of this optimal available expected degree of incorrectness, the *more irrational* the belief-system is. In general, the degree of irrationality of every available belief-system is measured by the *difference* between its expected degree of incorrectness and the optimal available expected degree of incorrectness.[10]

As I shall argue in the next section, for every thinker and every time t, there is always at least one belief-system b_t, which is available (at least in principle) to the thinker at t, such that b_t minimizes expected incorrectness according to the probability function that rationally should be guiding the thinker at t if she has this belief-system b_t at t. These belief-systems are the ones that it is maximally rational for the thinker to have at that time. In short, rational belief minimizes expected incorrectness.

This general account of rational belief will entail Joyce's dominance principle as a special case. If a belief-system b_1 is strongly dominated by an alternative b_2, that alternative b_2 will have a better – that is, lower – expected degree of incorrectness. So, the belief-system b_1 cannot minimize expected incorrectness. Since maximally rational belief-systems must minimize expected incorrectness, this strongly dominated belief-system b_1 cannot be maximally rational. In this way, the dominance principle can be explained on the basis of the more general and more fundamental principle that every perfectly rational belief-system must minimize expected incorrectness.

5. The structure of incorrectness

In the previous section, I proposed that a belief-system's degree of irrationality is determined by how its expected degree of incorrectness compares with that of the available alternatives. To develop this proposal, we will need to explore the nature of both (a) these degrees of incorrectness and (b) this

[9] I shall assume that there is a unique correct way of measuring these degrees of incorrectness – although beyond making a few remarks about this measure in the next section, I shall not be able to give a complete account of this measure here.

[10] For a related approach, which measures a credence function's "degree of incoherence" by its overall distance from the closest coherent function, see de Bona and Staffel (2017).

rational probability function – since these are the two elements in terms of which each belief-system's expected degree of incorrectness is defined.

As I have already implied, if the idea of a belief-system's expected degree of incorrectness even makes sense, then these degrees of incorrectness must be measurable on at least an interval scale. This measure can be represented by a scoring rule that assigns each belief-system an "incorrectness-score" – which can be represented by means of a real number. The structure of these degrees of incorrectness can be studied by exploring the features of this scoring rule.

I shall assume here that belief-systems can be modelled by credence functions – where each credence function assigns a real number between 0 and 1 to each of the propositions that the thinker has a doxastic attitude towards. The credence functions of some actual thinkers may be probabilistically incoherent (for example, a thinker may assign different credences to two propositions *A* and *B*, even though in fact *A* and *B* are logically equivalent); so not all of these credence functions will be probability functions.

In addition, I shall also assume that thinkers may be completely *attitudeless* towards certain propositions – indeed, they may even be attitudeless towards logically complex propositions like '$A \vee B$' even if they have attitudes towards the atomic propositions *A* and *B* out of which those complex propositions are composed. To capture this possibility, we must allow that these credence functions may be only partially defined: the set of propositions for which such a function is defined need not be a complete propositional algebra (that is, a field of propositions that is closed under operations like negation, disjunction, and the like). Probability functions, unlike credence functions, cannot be partially defined in this way: every probability function assigns a probability to every proposition in a complete algebra. Still, we may say that a credence function "coincides" with a probability function if and only if it is possible to *extend* the credence function into that probability function.[11]

Strictly speaking, to capture the ways in which actual thinkers' attitudes may be indeterminate, we should also allow that some belief-systems cannot be represented by a unique real-valued credence function, but only by a set of such functions. However, to keep things simple, I shall ignore this complication, and I shall pretend that every belief-system can modelled by a unique (though perhaps only partially defined) real-valued credence function.

The proposals that I have made so far already entail that this scoring rule must have certain features. First, it must have the feature that Joyce (2009, 279) calls "truth-directedness". Suppose that the truth values, true and false, are identified with 1 and 0 respectively. Then the incorrectness-score that a belief-system gets for each proposition *A* is determined purely by the *distance* between *A*'s truth value and the credence that the function that models the belief-system assigns to *A*; the smaller this distance, the better the score – and the best possible score is achieved when there is no distance at all between the credence and the truth value.

I have also proposed that the incorrectness-score for a whole belief-system is just a weighted sum of the incorrectness-scores that the belief-system gets for all the individual propositions that the thinker has doxastic attitudes towards. This implies that the scoring rule must have the feature that Joyce (2009, 271f.) calls "separability": the scoring rule has an additive structure of this kind.

There seem to be other features that it is intuitively plausible to ascribe to this scoring rule, such as "continuity" and "convexity", among others. But the most important feature for our purposes is what

[11] For the idea of extending a partially defined preference ranking into a complete ranking, see Joyce (1999, 103–6). The idea that I am invoking here is the analogue of that idea for belief-systems.

is known among statisticians as "propriety".[12] That is, the scoring rule must ensure that if each belief-system's expected incorrectness-score is calculated according to a given probability function P, then the belief-system that can be modelled by a credence function that coincides with P itself will always have an *optimal* expected incorrectness-score. If the scoring rule is not just "proper" but "strictly proper", then when calculated according to P, the belief-systems that can be modelled by credence functions that coincide with P will have the *uniquely best* expected incorrectness-score of all possible systems of credences.

For the rest of this discussion, I shall assume that this scoring rule is not just proper, but *strictly* proper.[13] Thus, when judged from the standpoint of a probability function P, the *only* belief-systems that have an optimal expected incorrectness-score are those that can be modelled by credence functions that coincide exactly with P itself. In the previous section, I proposed that the available belief-systems' degrees of irrationality are determined by their expected degrees of incorrectness according to the appropriate "rational probability" for the relevant thinker at the relevant time. This proposal, together with the strict propriety of the incorrectness scoring rule, implies that, for every thinker and every time, the maximally rational belief-systems for the thinker to have at that time will *always* be ones that can be modelled by credence functions that *coincide* with the appropriate "rational probability" function for that thinker at that time. It follows of course that every rationally optimal belief-system must be probabilistically coherent: it must match the particular probability function that counts as the rational probability for the thinker at the relevant time.

This assumption does not make the appeal to belief-systems' expected degree of incorrectness redundant. The notion of expected incorrectness is still necessary for explaining the degree of irrationality of rationally *sub-optimal* belief-systems. However, on these assumptions, the rationally optimal belief-systems will be those that coincide with the relevant rational probability function. So, on these assumptions, it is of great importance to understand exactly which probability function – out of all the infinitely many probability functions that exist – counts as the appropriate rational probability function of this kind. I shall consider this question in the next section.

6. The character of rational probability

Given that this rational probability function is a measure on the space of epistemically possible worlds that I described above, I have in effect already identified several features that this probability function must have. Specifically, this function must assign probability 1 to every logical and conceptual truth – including conceptual truths that (like 'If B, then Actually B') are not metaphysically necessary. Since metaphysical possibility and epistemic possibility come apart in this way, we should presumably also allow that truths that are metaphysically but not epistemically necessary (like 'Hesperus = Phosphorus') may have probabilities that are less than 1.

Why should this rational probability function have these features? The reason seems to be this. These logical and conceptual truths are in some way guaranteed to hold by the essential nature of our concepts, which are among the essential constituents of the mental states that we have – while truths that are metaphysically but not epistemically necessary may be guaranteed to hold by the essential nature of the relevant objects, properties, and relations, but are not in the same way guaranteed to hold by the essential nature of our mental states or their constituents.

[12] For discussions of the idea of a "proper" scoring rule, see for example Joyce (2009, 276), and Greaves and Wallace (2006).

[13] For some reasons in favour this assumption, see Wedgwood (2013a).

This suggests a general way to understand this rational probability function: it must capture everything that the essential nature of the thinker's mental states either guarantees, or at least makes likely, to be true. This understanding of this probability function fits naturally with an "internalist" view of rationality – that is, with the view that the attitudes that it is rational for a thinker to have at a given time are determined purely by the mental states and events that are present in the thinker's mind at that time, and not by any facts about the external world that could vary while these mental states and events remained unchanged.[14] At the same time, this understanding seems to address the main complaint that externalists have raised against internalism – namely, that if it is to be a genuine value, rationality must have a real (and not merely presumed) connection of some kind with the truth.[15]

One picturesque way of conceiving of this rational probability function is to imagine an *angel* perched inside the thinker's head – where the angel's advice to the thinker takes the form of this rational probability function. Unfortunately, this angel is uncertain about many empirical propositions about the world. However, the angel knows all relevant truths about the mental states and events that are present in the thinker's mind at the time; and she can assign probabilities to these empirical propositions by relying on what she knows about these mental states and events, together with what the essential nature of these mental states and events either guarantees or makes likely to be true. (Thus, for example, the angel knows all logical and conceptual truths involving concepts that the thinker possesses, among other things.)

As I explained in the last section, the proposals that I have made so far imply that at every time, every thinker rationally should have a system of beliefs that corresponds to the probability function that counts as the rational probability function for that thinker at that time (given that she has this system of beliefs). What this means, in effect, is that many of the great questions of epistemology become in effect questions about these rational probability functions.

Some features of the probability function that counts as the rational probability for a particular thinker at a particular time will vary between different thinkers and different times – even if these thinkers possess exactly the same concepts at these times. These features of these probability functions are in a sense *empirical*: they depend on contingent features of these thinkers' mental lives at these times, and not merely on the concepts that they possess or the capacities that are required for them to count as rational thinkers at all. Other features will be found in *every* probability function that counts as the rational probability of any thinker who possesses these concepts. These features are broadly *a priori*: they depend only on the concepts that these thinkers possess and the capacities that are presupposed by their being rational thinkers in the first place.

The fact that all logical and conceptual truths have probability 1 is presumably an *a priori* feature of these rational probability functions. But there may also be other such *a priori* features as well. For example, perhaps every one of these rational probability functions must obey the "Principal Principle" – so that, for example, if one of these probability functions assigns probability 1 to the proposition that a coin that is about to be tossed has a 0.5 chance of landing heads (and none of the other propositions that have probability 1 directly concern the outcome of the toss in an "inadmissible" way), then his function must also assign probability 0.5 to the proposition that the coin will land heads.[16]

[14] This is not the "accessibilist" version of internalism, but the "mentalist" version instead. I have discussed this kind of internalism elsewhere; see for example Wedgwood (2002).

[15] For such externalist criticisms of internalism, see for example Goldman (1999).

[16] For an illuminating discussion of the Principal Principle, see Meacham (2010).

There may also be certain universal principles governing the *empirical* features of these rational probability functions. Let us return to the image of this rational probability function as an angel perched inside your head, giving you advice on the basis of what she can work out from what she knows about the mental states and events inside your mind. This image suggests that one of the things that the angel will tell you about (even if you may often fail to take in the angel's advice) is everything that she knows about what is going on inside your mind. In other words, the rational probability function will incorporate a kind of *positive introspection principle*: if *A* is one of the relevant truths about the mental states that are currently present in your mind, then the rational probability function for you at this time must assign probability 1 to *A*.[17]

We can illustrate this introspection principle by considering the phenomenon of *self-undermining beliefs*. For example, consider Moore's paradox, which concerns the first-person present-tensed propositions that one could express by saying something of the form '*A* and I do not believe that *A*'. Clearly, such a proposition may be true. Moreover, nothing prevents this proposition from having high probability conditional on one's evidence. (Just suppose that you simultaneously receive messages from two fabulously reliable oracles – one testifying that you will never believe *A*, and the other testifying that *A* is true.) Still, it surely cannot be rational for you to believe this paradoxical proposition, even for an instant, because it should be obvious to you that if you believe this proposition, it is false.

Here is an explanation of why a perfectly rational thinker will never have such self-undermining beliefs, even for an instant. According to the introspection principle, for every thinker *x* and every time *t*, if *x* has a belief-system b_i at *t*, then the rational probability function for *x* at *t* must assign probability 1 to the proposition *that she has this belief-system b_i at t*. Clearly, no probability function can assign a high probability to '*A* and I don't believe that *A*' at the same time as assigning probability 1 to the proposition 'I believe that: *A* and I don't believe that *A*', at least not if it is effectively a conceptual truth that belief distributes over conjunction. (In that case, the probability function would also have to assign probability 1 to 'I believe that *A*' at the same time as assigning high probability to that proposition's negation 'I don't believe that *A*'.) According to my proposals, it is perfectly rational to have a certain credence in a proposition only if that credence matches the relevant rational probability of the proposition. Thus, it cannot be perfectly rational to have high credence in this Moore-paradoxical proposition.

This approach can also explain, not just self-undermining beliefs, but also *self-supporting* beliefs – such as the belief that one could express by saying 'I have at least one belief about my beliefs'. One might have no independent evidence supporting this belief, but it is obvious that if one believes this proposition, it will be true. So it seems that it is always rational to believe it. This can be explained on the basis of the fact that any probability function that assigns probability 1 to 'I believe that: I have at least one belief about my beliefs' must also assign probability 1 to 'I have at least one belief about my beliefs', given that it is a conceptual truth that the latter proposition is about one's beliefs, and so believing this proposition entails having at least one belief about one's beliefs.

In this way, the rational probability function that should be used to calculate each belief-system's expected degree of incorrectness (and so also, according to my proposals, each belief-system's degree

[17] If this strong "positive introspection" principle is acceptable, then arguably a corresponding "negative introspection" principle would also be acceptable: that is, if *B* is one of the relevant *false* propositions about what mental states you have at time *t*, then the rational probability function for you at *t* will assign probability 0 to *B*. These introspection principles may seem implausible, but they are not theses about what credence you will actually have – or even about the credences that you are realistically in a position to have. In the jargon of epistemologists (see Turri 2010), they are theses about *propositional*, not *doxastic*, justification – that is, about the attitudes that it is (ideally) rational for you to have, not about the attitudes that you will actually rationally have.

of irrationality) depends in part on which belief-system the thinker actually has. If the belief-system that the thinker has at *t* is b_i, then the probability function that determines the available belief-systems' expected degrees of incorrectness (which in turn determines b_i's degree of irrationality) must be a probability function that assigns probability 1 to the proposition that the thinker has this belief-system b_i at *t*.[18] It does not follow that this belief-system b_i is maximally rational – since even if this probability function assigns probability 1 to the proposition that the thinker has b_i, some other belief-system b_j might have a much better expected degree of incorrectness according to this probability function. (This will clearly be the case, for example, if b_i is probabilistically incoherent.)

In effect, this truth about what is present in the thinker's mind at the time can be thought of as at least part of the "evidence" that the thinker has at this time. Indeed, we can simply *identify* the "evidence" that the thinker has at this time with those propositions that (a) are assigned probability 1 by the rational probability function for that thinker at that time, but (b) unlike *a priori* truths are not assigned probability 1 by *all* rational probability functions for all thinkers and all times. On this conception of evidence, we may imagine that at each time the rational angel knows all the truths that constitute the thinker's evidence at the time – including all the truths about the beliefs and other mental states and events that are present in the thinker's mind at that time – and so the angel's advice is a probability function that gives all these evidence propositions probability 1. What perfect rationality requires of each thinker at each time is that she should have some belief-system that matches the probability function that would constitute the rational angel's advice given that she has that belief-system.

This idea could be developed in a number of ways. For example, some philosophers might suppose – perhaps along the lines that have been proposed by Timothy Williamson (2000, chap. 8) – that there is a special privileged *a priori* Ur-prior probability function, which *all* thinkers should start out from before any particular evidence is acquired. Given this supposition, it would be plausible to identify the rational probability function, for a given thinker at a given time, with the result of *conditionalizing* this *a priori* Ur-prior on the thinker's total evidence (including the truth about the mental states and events that are present in the thinker's mind at that time). In the picturesque terms that I have suggested, on this view, at every time the angel simply conditionalizes this *a priori* Ur-prior on the total evidence that the thinker has at that time.

However, many philosophers are sceptical about the idea of the special privileged Ur-prior probability function. Some of these philosophers might propose that so long as a belief-system meets certain basic constraints of synchronic coherence, "anything goes": every such synchronically coherent belief-system is perfectly rational. This is a kind of radical subjectivism about rational belief.[19] My idea could be developed so that it conforms to this subjectivist proposal – so long as it is one of the basic constraints of synchronic coherence that it is incoherent to have incorrect beliefs about the mental states and events that are currently present in one's mind. On this subjectivist proposal, the rational probability function for a thinker at a given time will always be a probability function that is as close as possible to the belief-system that the thinker actually has at that time. On this approach, the rational angel's advice at each time takes the form of a probability function that (a) meets all constraints of rational coherence, (b) assigns probability 1 to the evidence (including the truth about the mental states and events that are present in the thinker's mind at that time), and (c) is otherwise as close as possible to the thinker's actual belief-system at that time.

There is a third view that is intermediate between these two extremes – between the extreme rigorism of the *a priori* Ur-prior and the extreme permissivism of the radical subjectivist approach. This

[18] This feature of rational belief is in my view exactly paralleled by a corresponding feature of rational choice; for a discussion of this feature of rational choice, see Wedgwood (2013b).

[19] For example, van Fraassen (1984) might be taken as one proponent of such a radical subjectivist view.

intermediate view proposes that the thinker is rationally required to be guided by her *actual past* beliefs; but unlike with the theory of the special *a priori* Ur-prior, there is no reason to suppose that all rational thinkers must start out with the *same* set of ultimate priors. On this approach, the rational angel's advice at each time takes the form of a probability function that (a) meets all constraints of rational coherence, (b) assigns probability 1 to the evidence (including the truth about what is present in the thinker's mind at that time), and (c) is otherwise as close as possible to the belief-system that the thinker actually had *immediately before* that time.

In the next section, I shall explore this third proposal about the rational probability function in more detail. As I shall explain, this proposal introduces a new and deeper understanding of the notion of "diachronic rationality".

7. Diachronic rationality

I suggested above that one way to think of the distinction between synchronic and diachronic rationality is simply as marking a difference in the items whose rationality is being assessed – specifically, the difference between (a) statically enduring mental states (like belief-states) and (b) mental events (such as events in which we form or revise our beliefs).

There is, intuitively, a connection between these two kinds of rationality. Although one might form a belief in a proposition *A* and then immediately forget all about *A* in the next instant, the normal effect of forming a belief in *A* is precisely that one thereby acquires an enduring state of believing *A*. In general, the normal effect of each of these mental events is to have a certain constellation of static states, which we can view as the *output* of the mental event. The mental event itself occurs against the background of some *prior* mental states; we can view this background as the mental event's *input*. Thus, the mental event can be understood as a *transition* from this input to the relevant output.

An intuitively appealing account of the rationality of mental events of this kind is that such transitions are rational to the extent that they guarantee that their output will be no less rational than their input.[20] (So, in particular, a perfectly rational mental event will always map a perfectly rational input onto a perfectly rational output.) This account effectively defines the rationality of mental events in terms of the rationality of mental states.

There is also, however, a deeper question that is raised by the contemporary debates about diachronic rationality. This is the question of whether what I have called the "rational probability function", for a thinker at a given time, is determined purely by what is true of the thinker at that time, or whether it is also determined, at least in part, by something about the thinker's *past*.[21]

In the last section, I considered the radical subjectivist approach to rational belief. For these subjectivists, rationality is simply a matter of coherence between the attitudes that the thinker has at a single time. On this view, even if one's beliefs arbitrarily undergo a sudden radical revolution – as in the Biblical story of Paul's conversion on the road to Damascus (*Acts* 9: 3–9) – this need not be irrational in any way. However, just as we can make sense of coherence among the mental states that

[20] This does not yet amount to a precise general account of how to *measure* "the extent to which a transition guarantees that the output will be no less rational than the input". For cases where the thinker's evidence remains the same, one such measure is proposed by Staffel (2017); but it remains an open question how to develop such a measure for cases involving the acquisition (or loss) of evidence, where the input state is less than perfectly rational.

[21] This is the issue that is raised by the debate between synchronists like Hedden (2015) and their opponents like Podgorski (2016).

a thinker has at a time, we can also make sense of *coherence over time* – that is, coherence between the beliefs that a thinker has at one time and the beliefs that the thinker has at an immediately preceding time. Coherence over time is in effect a kind of *conservatism*: it involves minimizing the changes that one makes to one's belief-system in adjusting to new evidence over time. If rationality requires this sort of conservatism, this requirement would be diachronic in a deeper way.

One way to understand diachronic requirements of this deeper kind would be on the basis of the principle that I briefly sketched at the end of the previous section. This was the principle that if the thinker has a particular belief-system b_1 at a time t_1, then b_1's degree of irrationality is determined by how it compares with the available alternatives in terms of their expected degree of incorrectness according to a probability function P_1 that (a) meets all constraints of rational coherence, (b) assigns probability 1 to the evidence that the thinker has at that time (including the truth about the mental states, like the belief-system b_1, in the thinker's mind at t_1), and (c) is otherwise as close as possible to the belief-system b_0 that the thinker had at the *immediately preceding* time t_0.

How should we understand this idea of a probability function P_1's being "otherwise as close as possible" to the thinker's prior belief-system b_0? Consider the special case where the thinker's prior belief-system b_0 was perfectly probabilistically coherent, and so coincides with a probability function P_0. In this special case, the probability function that assigns probability 1 to the new evidence (including the truth about what is in the thinker's mind at t_1) but is otherwise "as close as possible" to the old belief-system b_0 *may* just be the result of *conditionalizing* P_0 on the evidence that the thinker has at t_1. However, as I shall explain in the next section, it need not *always* be the case that the rational probability function results from the thinker's prior belief-system by conditionalization in this way. If so, then conditionalization is at best an *approximation* to the truth about diachronic rationality. Unfortunately, however, I shall not be able to explore the conditions under which diachronic rationality does conform to conditionalization: that task will have to await another occasion.[22]

At all events, any version of this kind of conservatism about rational belief implies that every thinker is rationally required to have a certain kind of trust in her past beliefs. It is a good question why the thinker should trust her past beliefs in this way. Broadly speaking, it is analogous to other famous epistemological questions, such as why the thinker should trust her experiences or her episodic memories. Within our framework, these questions in effect become questions about the relevant "rational probability function". But the fundamental import of the questions remains the same.

I cannot attempt a full exploration of these questions here. But to fix ideas, I shall offer a conjecture about how they might be answered. According to this conjecture, the answer to this question involves three elements. The first element is the thesis that a tendency to maintain and continue relying on some of one's past beliefs is partially constitutive of being a thinker at all. The second element is the thesis that it is essential to all beliefs that they have some tendency to be rational, and as I have suggested, it is an essential and necessary feature of rationality that it has at least a certain weak connection to the truth. Finally, the third element is the point that both of the two extreme views of rational belief that were considered at the end of the last section are implausible. Contrary to the radical subjectivist view, our thinking would be objectionably arbitrary if there were nothing that guided us in responding to new evidence and new experiences. Contrary to the view that posits an *a priori* Ur-prior, there does not seem to be a special privileged Ur-prior that all thinkers should start

[22] This is one of several ways in which this conception of diachronic rationality differs from the familiar subjective Bayesian approach of the sort that was advocated by Jeffrey (2003). Besides not being committed to conditionalization's being the only rational method of belief revision, it is not restricted to cases where the thinker's prior beliefs are probabilistically coherent; it can allow for beliefs' being forgotten and evidence's being lost over time; and it may allow for further constraints of rational coherence (like the Principal Principle), in addition to probabilism itself.

out from. So there is nothing else in our minds that can reliably guide us in this way that has a stronger essential connection to the truth than our past beliefs.

Together, these three points may provide an explanation of why we should trust our past beliefs in this way – specifically, an explanation that respects both the internalist insight that rationality supervenes purely on the mental states and events in the thinker's mind, and also the point (which externalists rightly insist on) that rationality must have some real and not merely imagined connection to the truth. In this way, the proposals that I have suggested so far can provide a sketch of how the alethic-value-based conception of epistemology might account for diachronic requirements of rationality.[23]

8. Self-supporting and self-undermining beliefs

In this last section, I shall explain how the picture that I have sketched above can deal with *self-supporting* and *self-undermining* beliefs, which some philosophers have taken to provide challenging problems for this approach. My solution to these problems will rely heavily on the idea that the rational probability function satisfies the positive introspection principle that I described above.

My proposals imply that a perfectly rational belief-system can always be modelled by a credence function that coincides with the relevant rational probability function. As I understand this, a credence function can "coincide" with a probability function even if the credence function is only partially defined. For the rest of this discussion, however, I shall ignore this complication. I shall assume that the thinker does have attitudes towards all the propositions that feature in the cases that we shall consider. This means that for our purposes, we need not worry about the distinction between a credence function (which may be only partially defined) and a probability function (which is defined for every proposition in the relevant algebra). In effect, we will work with the simplifying assumption that every rational belief-system can be modelled by a credence function that is identical to a probability function.

My proposals about the rational probability function also incorporate a positive introspection principle, according to which the rational probability function must assign probability 1 to the truth about the relevant mental states and events that are present in the thinker's mind at the relevant time. Taken together, these proposals imply that, at least in the relevant cases, there are some significant limits on what it is rational to believe. In particular, as has been shown by Andy Egan and Adam Elga (2005), if the thinker correctly introspects her level of credence in a proposition *A*, it cannot be rational for her to have too high a level of confidence in the biconditional proposition that she could express by saying '*A* is true if and only if it is not the case that my credence in $A \geq 0.5$'. As Egan and Elga put it, you cannot rationally believe that you are an "anti-expert" about *A*.

Suppose that the thinker has credence 1 in the biconditional that she could express by saying '*A* is true if and only if it is not the case that my credence in $A \geq 0.5$'. Now, either her credence in $A \geq 0.5$ or it is not. Suppose that her credence in $A \geq 0.5$. Then she will have a maximally confident introspective belief that her credence in $A \geq 0.5$; and given her belief in the biconditional, this introspective belief will commit her to having a credence of 0 in *A*. Alternatively, suppose that it is not the case that her credence in $A \geq 0.5$. Then the thinker will introspect that fact about her beliefs, and together with the biconditional, she will be committed to having a credence of 1 in *A*. So, either way, if she has too much confidence in this biconditional, introspection guarantees probabilistic incoherence.

[23] This sketch of an explanation of why it is rational for us to be guided by our past beliefs is modelled on the account of why it is rational for us to trust our sensory experiences that I gave elsewhere (Wedgwood 2011).

Some philosophers have tried to exploit this point to raise objections to probabilism, or to the idea that a perfectly rational belief-system must minimize expected incorrectness. Thus, for example, Carr (forthcoming) imagines that a "perfectly reliable … teacher" might tell you that A is true if and only if it is not the case that your credence in $A \geq 0.5$. But even if a perfectly reliable teacher tells you this, according to my proposals, it cannot be rational for you to believe what this teacher says. In effect, if your teacher tells you this, your only rational option is to doubt your teacher's reliability.

We need not deny that you might learn that you are *going to be* an anti-expert about some proposition in the *future*. You might rationally come to believe at t_0 that A is true if and only if it is not the case that the credence in A that you will have at a later time $t_1 \geq 0.5$. But when t_1 comes, you cannot rationally retain this belief by believing the proposition that you could express by uttering the biconditional 'A is true if and only if it is not the case that the credence that I *now* have in $A \geq 0.5$'. Admittedly, there may be no way of explaining how you rationally lose your confidence in this biconditional at t_1 by appeal to anything like conditionalization. If so, however, what this would show is that conditionalization is only an approximation to the full story about diachronic rationality – an approximation that gets it right in wide range of cases, perhaps, but not in tricky cases of this sort.[24]

Michael Caie (2013) has focused on cases of self-referential propositions – like the propositions that you might express by saying something like 'I do not believe *this very proposition*'. More precisely, he focuses on propositions that you might express by saying 'It is not the case that my confidence in the truth of *this very proposition* ≥ 0.5'. Caie assumes that as a matter of logic or analytic truth, this proposition is true if and only if it is not the case that your confidence in this proposition ≥ 0.5.

However, I do not see why we should accept that this is a matter or logic or analytic truth. Every proposition that you are capable of even considering or having any attitudes towards, I propose, is such that there is some doxastic attitude that it is perfectly rational for you to have towards it.[25] Given the conception of rational belief that I am sketching here, there could not be any perfectly rational attitude for you to have towards a proposition A for which it was an analytic truth that A is true if and only if it is not the case that your confidence in $A \geq 0.5$. So I must conclude that it is not possible for you even to consider or to have any attitudes towards any such proposition. Whatever the sentence 'I do not believe this very proposition' expresses, as uttered on your lips, it cannot be a proposition of this sort.[26]

I shall close by considering a number of cases that have been explored by Hilary Greaves (2013), which she takes to constitute problems for the general idea of an alethic-value-based epistemology of the sort that we have been discussing here. Officially, every one of these cases concerns *synchronic* rationality, since these cases are stipulated to be cases in which the thinker acquires no new evidence. Greaves is interested in exploring a version of this value-based approach that she takes to be analogous to "practical decision theory" – that is, to the theory of rational choice. According to this version of the approach, having a certain belief-system is construed an "epistemic act", where the rationality of such acts is explained by the kind of practical decision theory that she has in mind. According to this kind of decision theory, the rationality of an act is explained by a set of credences as well as by an appropriate value. So, Greaves specifies each of the cases that she considers by means

[24] For a compelling argument for the conclusion that conditionalization is only an approximation to the full story about diachronic rationality, see Arntzenius (2003).

[25] The underlying idea is that propositions of the relevant kind are Fregean *Gedanken*, which as I have argued elsewhere (Wedgwood 2007, Chap. 7), are essentially individuated by the conditions under which it is rational to take doxastic attitudes towards them.

[26] This approach to these self-referential cases is due to Prior (1961). I am indebted here to conversation with my colleague Andrew Bacon.

of what she calls an "initial" set of credences. In each case, certain belief-systems or sets of credences turn out to maximize some kind of expected value according to those initial credences; she labels each set of credences that maximizes this kind of expected value the thinker's "final" credences.

This interpretation of the value-based framework has a number of features that seem somewhat strange, at least to me. First, it is odd to compare the statically enduring state of having a certain belief-system to an *act*: acts are naturally thought of as events, rather than statically enduring states; further, acts typically involve part of what occurs in the agent's environment, as well as what occurs in the agent's mind. On the practical side, the item that seems most analogous to a belief-system is not an act or even a decision but rather a collection of *plans* or *intentions*, which like a belief-system is a statically enduring state. (In formal presentations of decision theory, the item whose rationality is being explained is typically identified with a *system of preferences*, but if we stick more closely to ordinary folk-psychological thought, it seems that the practical mental state whose rationality is in question is typically a collection of plans and intentions.) This interpretation of practical decision theory as fundamentally concerned with the rationality of collections of intentions also suits the fact that, strictly speaking, decision theory is a purely *synchronic* theory, specifying the relations of coherence that rationality requires between the preferences, credences, and intentions that the agent has at any given time; as a synchronic theory, it is more easily understood as primarily concerned with statically enduring mental states than with mental events or transitions.

Secondly, it is hard to understand exactly what role the "initial credences" are supposed to play in the cases that Greaves describes. Since the thinker acquires no new evidence, surely the thinker should just persist with the initial credences that she already has. (Even if we understand these "initial credences" as corresponding to what I have called the relevant "rational probability" function, it seems that the thinker should just have a set of credences that matches this probability function.) What Greaves (2013, 936) herself says is the following: "In other words, the agent's awareness of certain facts – the facts given in the case-specification – gives rise to rationality constraints on her credences; the 'initial' credence function is one respecting these rationality constraints." But if (because of the thinker's "awareness of certain facts") the thinker really is rationally required to have a credence function that respects certain constraints, then it cannot be true that she is rationally permitted at the same time to have a "final" credence function that does not respect these constraints. So the role of these "initial" credences remains obscure.

Still, Greaves's description of the cases seems intuitively to make sense. It seems to me that when we have an impression that these cases make sense, we are tacitly forgetting the stipulation that these cases do not involve the acquisition of evidence. We are supposing instead that the thinker *learns* that the case that she is in has the specified features, and we then inquire into how it is rational for the thinker to *respond* to acquiring this information. For example, consider what Greaves (2013, 916) calls the "Leap" case:

> Bob stands on the brink of a chasm, summoning up the courage to try and leap across it. Confidence helps him in such situations: specifically, for any value of *x* between 0 and 1, if Bob attempted to leap across the chasm while having degree of belief *x* that he would succeed, his chance of success would then be *x*. What credence in success is it epistemically rational for Bob to have?

Intuitively, this case makes perfect sense. But this, I suggest, is because we interpret it as a case in which Bob *learns* that his chance of success in leaping across the chasm corresponds exactly to his degree of belief in the proposition that if he tried to leap across, he would succeed. The question that we want to answer about this case is: how (if at all) should Bob revise his degree of belief in this proposition in response to learning this?

On this interpretation of this case, it is clear what the theory that I have proposed will say. It will say that *whatever* credence Bob might have in this proposition, that credence will be perfectly rational.

This is because every such credence *x* forms part of a belief-system b_x such that *x* is the probability that is assigned to the proposition in question by the probability function that meets all constraints of rational coherence (including the Principal Principle), assigns probability 1 to all of Bob's evidence (including the proposition that Bob has this belief-system b_x), and is otherwise as close as possible to Bob's prior belief-system.

Admittedly, it is tempting to say that in this case, Bob should have a more extreme credence, such as either 0 or 1. However, this inclination is probably explained by the difficulty in separating an austerely epistemic use of 'should' from a more *practical* use. In a more practical sense, these extreme credences have something to be said for them that the more intermediate credences lack: these extreme credences will lead to definite safe decisions, either to leap across the chasm, or to step back and find another route forward, while the more intermediate credences will lead to a more difficult and alarming decision. However, if we set these practical considerations aside, there seems to me nothing strictly epistemic to be said in favour of any of these credences over any of the others. (Imagine that Bob is not trying to leap across the chasm, but is just inquiring in a spirit of purely idle curiosity about whether he would succeed if he tried. Then it does not seem obvious that it is less rational for him to have an intermediate credence like 0.5 in this proposition rather than an extreme credence like 0 or 1.)

With all of Greaves's other examples, the theory that I have proposed will straightforwardly agree with the verdicts that she regards as intuitively plausible. For example, consider Greaves's (2013, 915f.) "Promotion" case:

> Alice is up for promotion. Her boss, however, is a deeply insecure type: he is more likely to promote Alice if she comes across as lacking in confidence. Furthermore, Alice is useless at play-acting, so she will come across that way iff she really does have a low degree of belief that she's going to get the promotion. Specifically, the chance of her getting the promotion will be $(1 - x)$, where *x* is whatever degree of belief she chooses to have in the proposition *P* that she will be promoted.

Here, there is a uniquely rational credence for Alice to have – 0.5. This is the only available credence that matches the rational probability function that should be guiding her given that she has all this evidence (including the fact that she has this very credence).

In all these cases, however, it is not really the value-based framework itself that explains why my approach yields the intuitively correct verdicts. What explains this is the character of the rational probability function. Specifically, it is explained by the following two features of this probability function:

i. First, for every thinker *x*, time *t*, and belief-system b_t, it is perfectly rational for *x* to have b_t at *t* if and only if b_t matches the probability function that counts as the rational probability for *x* at *t*, given that the thinker *x* has this belief-system b_t at *t*.
ii. Secondly, whenever the thinker's evidence changes, this probability function also changes to the probability function that (a) conforms to all constraints of rational credence, (b) assigns probability 1 to all the thinker's evidence (including the truth about the new belief-system that the thinker now has), and (c) is otherwise as close as possible to the belief-system that the thinker had at the immediately preceding time.

As condition (b) of this second point (ii) makes clear, these features of the rational probability function include the positive introspection principle that I have described. In general, the explanation of these features of the rational probability function relies on the internalist-but-still-truth-connected conception of rationality that I alluded to above; it does not directly rely on the idea that rational belief-systems must minimize expected incorrectness.

In this way, my proposed interpretation of the connection between epistemic rationality and minimizing expected incorrectness (or optimizing expected alethic value) is not, as Greaves (2013, 926) puts it, "ambitious". That is, it does not attempt to derive *all* of the distinctive features of rational belief from this idea of the connection between rationality and alethic value. But I suggest that the goal of finding such an ambitious interpretation is chimerical. Once we abandon the attempt to give such an ambitious interpretation, we can understand how rationality is connected to correctness – that is, to the idea of alethic value – without sacrificing the plausibility of our account of rationality.[27]

References

Åqvist, Lennart (1984). "Deontic Logic", in Dov Gabbay, ed., *Handbook of Philosophical Logic* (Dordrecht: Reidel), 605–714.

Arntzenius, Frank (2003). "Some problems for conditionalization and reflection", *Journal of Philosophy* 100 (7): 356-370.

Berker, Selim (2013). "The Rejection of Epistemic Consequentialism", *Philosophical Issues* 23: 363-387.

Broome, John (1991). "'Utility'", *Economics and Philosophy* 7 (1): 1-12.

Caie, Michael (2013). "Rational Probabilistic Incoherence", *Philosophical Review* 122 (4): 527-575.

Carr, Jennifer (2015). "Epistemic Expansions", *Res Philosophica* 92 (2): 217-236.

——— (forthcoming). "Epistemic Utility Theory and the Aim of Belief".

de Bona, Glauber, and Staffel, Julia (2017). "Graded Incoherence for Accuracy-Firsters", *Philosophy of Science* 84 (April): 189–213.

Egan, Andy, and Elga, Adam (2015). "I can't believe I'm stupid", *Philosophical Perspectives* 19 (1): 77–93.

Firth, Roderick (1981). "Epistemic Merit, Intrinsic and Instrumental", *Proceedings and Addresses of the American Philosophical Association* 55: 5–23.

van Fraassen, Bas (1984) "Belief and the Will", *Journal of Philosophy* 81 (5): 235–256.

Goldman, Alvin (1999). "Internalism Exposed", *Journal of Philosophy* 96 (6): 271–293.

Greaves, Hilary, and Wallace, David (2006). "Justifying conditionalization: Conditionalization maximizes expected epistemic utility", *Mind* 115 (459): 607-632.

[27] Earlier versions of this paper were presented at the University of Kent, the University of Konstanz, and the University of Michigan. I am grateful to the members of those audiences – and especially to my commentator in Kent, Richard Pettigrew, and my commentator in Michigan, Daniel Drucker – and also to Jeffrey Dunn, one of the editors of this volume, for helpful comments. Finally, I am also indebted to Catrin Campbell-Moore, Jason Konek, Ben Levinstein, and Julia Staffel, who had an online discussion of this paper – and especially to Julia Staffel, who gave me an extreme useful summary of the group's comments.

Greaves, Hilary (2013). “Epistemic Decision Theory”, *Mind* 122 (488): 915-952.

Hansson, Bengt (1968). “Choice Structures and Preference Relations”, *Synthese* 18: 443–58.

Krantz, David H., Luce, R. Duncan, Suppes, Patrick, and Tversky, Amos (1971). *Foundations of Measurement* I*: Additive and Polynomial Representations* (London: Academic Press).

Meacham, C. J. G. (2010). “Two Mistakes Regarding the Principal Principle”, *British Journal for the Philosophy of Science* 61 (2): 407–431.

Hedden, Brian (2015). “Time-Slice Rationality”, *Mind* 124 (494): 449-491.

Jeffrey, Richard C. (2003). *Subjective Probability: The Real Thing* (Cambridge: Cambridge University Press).

Joyce, James M. (1998). “A Nonpragmatic Vindication of Probabilism”, *Philosophy of Science* 65 (4): 575-603.

——— (1999). *Foundations of Causal Decision Theory* (Cambridge: Cambridge University Press).

——— (2009). “Accuracy and Coherence: Prospects for an Alethic Epistemology of Partial Belief”, in Franz Huber and Christoph Schmidt-Petri, ed., *Degrees of Belief*, Synthese Library Vol. 342 (Berlin: Springer).

Kratzer, Angelika (2012). *Modals and Conditionals: New and Revised Perspectives* (Oxford: Oxford University Press).

Lewis, David (1972). *Counterfactuals* (Oxford: Blackwell).

Pettigrew, Richard (2013). “Introducing … Epistemic Utility Theory”, *The Reasoner* 7 (1): 10–11.

Pettit, Philip (1991). “Consequentialism”, in Peter Singer, ed., *A Companion to Ethics* (Oxford: Blackwell).

Podgorski, Abelard (2016). “A Reply to the Synchronist”, *Mind* 125 (499): 859–871.

Prior, A. N. (1961). “On a family of paradoxes”, *Notre Dame Journal of Formal Logic* 2 (1): 16–32.

Rawls, John (1971). *A Theory of Justice* (Cambridge, Massachusetts: Harvard University Press).

Staffel, Julia (2017). “Should I pretend I’m perfect?” *Res Philosophica* 94 (2): 301–324.

Thomson, Judith (2009). *Normativity* (Chicago, IL: Open Court).

Turri, John (2010). “On the relationship between propositional and doxastic justification”, *Philosophy and Phenomenological Research* 80 (2): 312–326.

Wedgwood, Ralph (2002). “Internalism Explained”, *Philosophy and Phenomenological Research* 65: 349–369.

——— (2007). *The Nature of Normativity* (Oxford: Clarendon Press).

——— (2009). “The ‘Good’ and the ‘Right’ Revisited”, *Philosophical Perspectives* 23: 499–519.

——— (2011). “Primitively Rational Belief-Forming Processes”, in *Reasons for Belief*, ed. Andrew Reisner and Asbjørn Steglich-Petersen (Cambridge: Cambridge University Press): 180–200.

——— (2013a). “Doxastic Correctness”, *Proceedings of the Aristotelian Society*, Supp. Vol. 87: 38–54.

——— (2013b). “Gandalf’s Solution to the Newcomb Problem”, *Synthese* 190 (14): 2643–2675.

——— (2014). “Rationality as a Virtue”, *Analytic Philosophy* 55 (4): 319-338.

Williamson, Timothy (2000). *Knowledge and its Limits* (Oxford: Clarendon Press).