0. ‘Rationality’ and ‘entitlement’

Many terms can be used to evaluate beliefs, belief-forming processes, and other similar mental states and processes of reasoning. Such mental states and processes can be called “right” or “wrong”, “correct” or “incorrect”, “reasonable” or “unreasonable”, “justified” or “unjustified”, and so on. In recent years, ever since Tyler Burge gave his John Locke lectures in Oxford in 1993, many philosophers evaluate mental states or processes of reasoning by saying that they are mental states that the thinker in question is “entitled” to have, or processes of reasoning that the thinker is “entitled” to perform.1 As these philosophers would say, a central part of epistemology is concerned with investigating the conditions under which thinkers are in this sense “entitled” to various beliefs and processes of belief-formation or belief-revision.

Other epistemologists favour the word ‘rationality’. This term is particularly popular among formal epistemologists – such as those in the Bayesian tradition, which ascribes a starring role to the mathematical notion of probability.2 According to this alternative tradition, the central questions of epistemology concern the conditions under which our beliefs, and the processes by which we form and revise our beliefs, count as “rational”. As I shall explain in this section, this difference between the two epistemological traditions is fundamentally terminological rather than conceptual. The terms ‘rationality’ and ‘entitlement’ are used in different epistemological traditions, but they express the same range of concepts.

Both terms, when they are used by epistemologists in these ways, seem to express broadly normative concepts. The term ‘entitlement’ seems normally to express a normative concept. Indeed, the natural home of this term seems to be within systems of institutional rules: for example, if you are “entitled”, under the rules of the university that you work for, to go on sabbatical leave for a certain period of time, then according to those rules, it is permissible for you to go on sabbatical leave for that period; and if you communicate your intention to go on leave, then the university will normally be obliged to allow you to take that period of sabbatical leave. Similarly, the term ‘rational’ can also be used to express a normative concept – specifically, a concept that refers to the proper use of one’s faculties of thinking and reasoning. To say that you “rationally believe” a certain proposition \( p \) is to say that in believing \( p \), you are using these faculties properly in the relevant sense; and to say that “it is rational” for you to believe a certain proposition \( q \) is to say that there is a possible process of reasoning, consisting of the proper use of these faculties, that leads from your current state of mind to your believing \( q \).

Among epistemologists who use the term ‘entitlement’ it is widely supposed that “entitlement” differs in an important way from “justification”. In fact, different philosophers have proposed several different contrasts between justification and entitlement. For example, Crispin Wright (2004, 167) says that entitlement is “a kind of rational warrant” for accepting a proposition that is not to be identified with “having evidence for its truth”. Yuval Avnur (2012, 299) gives a somewhat different characterization when he says:

---

1 See especially Burge (1993 and 2003) for his most influential discussions of entitlement. Another philosopher who articulates the central questions of epistemology in terms of entitlement is Peacocke (2003).
2 For a classic statement of this Bayesian approach, see Jeffrey (2004).
Entitlement is a positive epistemic status that can apply to either a belief or a belief-forming procedure. Having entitlement does not require any independent evidence or reason for the belief in the first case or for believing that the belief-forming procedure is reliable in the second case. Let us stipulate that evidence that one’s faculties are reliable is independent if and only if it was not produced by (or accessed via) those very faculties.

For both Avnur and Wright, then, entitlement differs from other kinds of warrant in that entitlement does not require “evidence”.

It is clear that the overwhelming majority of formal epistemologists would agree that there are many cases where it is rational for a thinker to have a certain level of confidence in a proposition \( p \), and what makes it the case that this is rational is not any “evidence” that the agent possesses for or against this proposition \( p \). For example, according to subjective Bayesians like Richard Jeffrey (2004), one’s “prior” credences are automatically rational, so long as they are probabilistically coherent; the rationality of these prior credences does not depend on evidence that the thinker has for the truth the propositions in question.

So-called objective Bayesians reject the idea that our prior credences are automatically rational in this way; they may even say that there is only one possible set of prior credences that count as rational (such as the credences that conform to some version of the notorious “principle of indifference”). However, even for these objective Bayesians, it cannot be the evidence that the thinker has that makes these prior credences rational, since these are the credences that it is rational for the thinker to have in advance of all evidence whatsoever. According to Avnur and Wright’s understanding of “entitlement” and the formal epistemologists’ understanding of “rationality”, then, it seems clear that the beliefs that a thinker is “entitled to” are precisely those beliefs that it is “rational” for the thinker to hold.

A different contrast between entitlement and other kinds of warrant is suggested by Burge. According to Burge (2003, 504), the key difference between entitlement and other kinds of warrant is that entitlement does not require that the thinker has “conceptual access” to the entitlement: as he says, “individuals can be epistemically entitled to a belief … without having the concepts needed to understand or even think the entitlement.”

While Burge’s formulations could be read in more than one way, it seems clear that whatever features he is ascribing to entitlement here, they are features that Bayesians and other formal epistemologists would equally ascribe to rationality. According to proponents of Bayesianism, the rational thinker need not have the concepts of probabilistic coherence, or of a warrant, or of a reason, or anything of that sort. She need not have the concept of evidence or even the concept of belief; she need not even have the concept of experience or perception, or the concept of a belief’s being incoherent. According to the subjective Bayesians, all that is required is that the thinker’s beliefs must actually be probabilistically coherent, and they must evolve in response to evidence by means of conditionalization; objective Bayesians would impose further conditions, but none of the conditions that have been proposed require that the thinker must possess any of these concepts. Here too, then, the notions expressed by the term ‘entitlement’ seem to coincide with those that are expressed by ‘rationality’.

There is, admittedly, one significant difference between the tradition that speaks of “entitlement” and the tradition that favours the term ‘rationality’. The former tradition is concerned exclusively with

---

3 For an example of an objective Bayesian approach, see Leitgeb and Pettigrew (2010).
4 For an illuminating discussion of how to interpret Burge’s work, see Casullo (2007).
epistemology – with the conditions under which we are entitled to believe various propositions. The latter tradition is just as much concerned with rational choices and rational preferences as with rational beliefs. Offhand, it seems that we can apply the same terms, in the same range of senses, to both beliefs and choices; for example, everyday language speaks both of reasons for belief and of reasons for choice, with no sign that the term ‘reason’ changes its meaning between the two contexts. In what follows, I shall aim to speak about rationality in full generality, encompassing both rational belief and rational belief.

For this reason, I shall from now on stop explicitly discussing “entitlement”; I shall speak solely of “rationality” instead. As I have just argued, this is fundamentally a change in terminology, but not a change in the basic subject.

1. A problem about rationality

As I have already stated, ‘rational’ seems often to express a normative concept – as I put it, a concept that has to do with the proper use of one’s reasoning and thinking faculties. When it is used in this way, ‘rationality’ seems to have intimate connections with several other paradigmatically normative terms. The notion of what is “rationally required” of you at a time t is a kind of ‘ought’ (focused on your cognitive situation at t): if you are rationally required not to believe a certain proposition p at t, then in some sense, you ought not to believe p at t.

Moreover, in a broad sense, rationality is a kind of virtue that can be exemplified by the thinking or reasoning that you perform at a certain time – while irrationality is the corresponding vice. That is, it seems to be a conceptual truth, flowing form the nature of the concept that is expressed by this use of ‘rationality’, that rationality is good feature of reasoning and thinking, while irrationality is a correspondingly bad feature of reasoning and thinking.

Indeed, rationality is arguably a particularly fundamental virtue of thought, in the sense that all thinkers whose thinking is intelligible at all must have at least some disposition to reason in more-or-less rational ways. In general, we typically find it intelligible when thinkers respond to their evidence in a broadly rational way, and we find irrationality more puzzling and harder to make sense of.

At the same time, rationality seems to be a normative notion of a distinctive kind. When we assess a mental state like a belief as right or wrong, or as correct or incorrect, we are typically assessing it on the basis of its relation to the external world. By contrast, when we assess a mental state like a belief as rational or irrational, we are assessing it purely on the basis of its relation to what is going in the thinker’s mind. That is to say, a certain modest kind of internalism seems to be true about rationality. This is not the “accessibilist” kind of internalism, according to which rationality depends on what the thinker has some kind of introspective or reflective “access” to; it is the more austere “mentalist” kind of internalism, according to which rationality supervenes purely on the mental states and mental events and processes that are present in the relevant thinker’s mind.

To avoid misunderstanding, I should emphasize that this sort of mentalist internalism about rationality is compatible with an anti-individualist conception of the mental states themselves. According to this anti-individualist conception, which mental states you are in depends, in part, on your relations to

---

5 For a highly influential development of this idea, see Davidson (1980, Essays 11 and 12; and 2001, Essays 9 and 10).
6 I have defended this sort of internalism elsewhere; see especially Wedgwood (2002 and 2006).
7 For the distinction between these two forms of “internalism”, see Conee and Feldman (2001).
your environment, and not just on what is going on inside your head. This is compatible with the claim that whether or not your beliefs are rational depends purely on the mental states that you are in, and not on any features of your environment that could vary while those mental states remained unchanged.

In short, it seems plausible to claim that the concept of “rationality” is both (i) normative and (ii) internalist in this way. This claim could be challenged, but I shall assume for the sake of the present discussion that it is true. If it is true, however, it immediately creates an obvious intuitive problem.

According to the kind of internalism that I am assuming here, satisfying a requirement of rationality simply consists in exemplifying a certain pattern among your mental states and mental events. But how can we reconcile this point with the assumptions that “rationality” is a normative concept, and that rationality is a fundamental virtue of thought? After all, what is so important about exemplifying this internal pattern of mental states and events? How can this mental pattern be a requirement of a fundamental virtue of thought?

This problem seems to call for a kind of explanation of how there can be a fundamental virtue of thought that requires exemplifying this internal mental pattern. More precisely, what is called for is an explanation of the requirements of rationality that makes it clear how exactly these requirements matter. As I shall put it for short, it calls for an explanation of why rationality matters.

There are three possible responses to this problem:

a. There is no further explanation of why this requirement of rationality matters: it is simply a brute primitive fact about the normative domain.

b. There is an explanation of why this requirement of rationality matters; and every such explanation appeals only to internal facts about what is going on in the thinker’s mind.

c. There is an explanation of why this requirement of rationality matters; and the explanation appeals, at least in part, to something that is external to the thinker’s mind.

According to the first response (a), there is no absolutely explanation that can be given of why this requirement of rationality matters. As I see it, this response has two fundamental problems.

First, it seems that the requirements of rationality are a rather miscellaneous assortment. If there is no explanation of why any requirements of rationality matter, it is doubtful whether we could explain what unifies all these different requirements (since any such unifying explanation would surely explain why at least some of these rational requirements matter). But there must surely be something that unifies all the requirements of rationality: they surely cannot just be a random heap of requirements. So, there must be an explanation of why at least some of the requirements of rationality matter; and if some of these requirements can be explained, our account would succeed in unifying the requirements of rationality even more cogently if all requirements of rationality could be explained in fundamentally the same way.

Secondly, this response to the problem seems to imply that it matters in itself, purely for its own sake, whether or not one’s thinking exemplifies the internal pattern that is in question. But why should such a thing matter purely for its own sake? Perhaps thinking that does not meet these internal requirements of rationality forms a less pretty mental pattern than thinking that does meet these requirements. But surely that can’t explain why we “ought” (in the particularly fundamental sense at issue) always to conform to this requirement. In short, it seems intrinsically implausible to claim that there are serious normative requirements of this kind for which no deeper or more illuminating
explanation can be given. So, I recommend, we should at least provisionally set this first response aside and explore the remaining two responses.

2. Dutch books?

Many decision theorists and formal epistemologists are intensely interested in the quest for an explanation of these rational requirements. One of the most popular approaches among these formal epistemologists involves appealing to “Dutch book arguments”.

These arguments start out from the following point. Suppose that there is some logical truth — say, $p$ — of which you are not absolutely certain. Then a clever bookmaker would be able to sell you a bet that would give you a monetary gain only if $p$ turns out to be false. For example, if you have a 5% degree of belief that $p$ is false, and a 95% degree of belief that $p$ is true, you would be willing to pay up to $1 for a bet that will pay you $20 if $p$ is false, and nothing if $p$ is true. It would then be guaranteed by logic alone that you will lose $1 to this bookmaker whatever happens. Any set of bets including this bet is a so-called “Dutch book” — a set of bets that taken together guarantee a certain loss, no matter how things turn out. On the other hand, if you are absolutely certain of this logical truth $p$, then the only bets that you will accept on $p$ are bets in which you lose absolutely nothing if $p$ is true — and none of those bets will guarantee that you will lose money no matter what happens.

The Dutch book arguments rely on theorems that generalize this point. The most famous theorem that formal epistemologists have proved is that whenever your credences are probabilistically incoherent, you will be willing to accept a Dutch book of this kind — whereas if your credences are probabilistically coherent, you will not be vulnerable to a Dutch book in this way. According to the argument, this theorem provides an explanation of why rationality requires each of us to maintain probabilistically coherence in our credences.  

Still, it is not clear how exactly this theorem explains this alleged requirement of rationality. The argument might be that there is something bad about losing money, and probabilistic coherence is rationally required because of the relationship between probabilistic incoherence and losing money. But then this approach is a version of the third response to our problem: it sets out to explain rational requirements by relating the violation of such requirements to a kind of event that may occur in the external world — namely, the event of one’s losing money. I shall consider such external explanations of rational requirements in the second half of this paper (starting in Section 4).

However, there is also a second way of understanding the appeal to this Dutch book theorem. As David Lewis (1999, 404–5) put it:

Note also that the point of any Dutch book argument is not that it would be imprudent to run the risk that some sneaky Dutchman will come and drain your pockets. After all, there aren’t so many sneaky Dutchmen around; and anyway, if ever you see one coming, you can refuse to do business with him. Rather, the point is that if you are vulnerable to a Dutch book … you hold two contradictory opinions about the expected value of the very same transaction. To hold contradictory opinions may or may not be risky, but it is in any case irrational.

This alternative interpretation of how the Dutch book theorem explains requirements of rationality makes this explanation a version of the second response to our problem: it bases its explanation purely on internal facts about the thinker’s mind.

---

8 For an illuminating discussion of these Dutch book arguments, see Hájek (2005).
It is clear, however, that this second response to our problem has a troubling feature. On this interpretation, the explanation simply presupposes that it is irrational (as Lewis puts it) “to hold contradictory opinions”. So this approach cannot explain why it is irrational to have contradictory opinions.

Thus, this approach can at best solve the first problem (i) for the first solution (a) to our problem – at best, it could give an account that unifies the various different requirements of rationality; but it cannot solve the second problem (ii) – it cannot explain why these requirements of rationality are more than just a pretty internal pattern of mental states. This gives us a reason to continue our search for an explanation of the requirements of rationality.

3. Constitutive truths?

Some philosophers have been tempted by the idea that the basic principles of rationality are in some way “constitutive” of the very nature of the relevant mental states. One version of this idea concerns the concepts that we possess: according to this version of the idea, for every concept that you possess, you must be disposed to use the concept in some of the ways in which it is rational to use it. For example, perhaps you could not possess a logical concept, like the concept that is expressed by “if”, unless you had some disposition to use it in accordance with the logically valid rules of inference involving this concept, such as the rule of modus ponens. If someone had no disposition to accept the appropriate inferences involving a certain concept, then that concept cannot be the concept ‘if’.

Another version of this idea concerns the types of attitudes that we are capable of. Perhaps we would not count as having beliefs, or as making choices or the like, unless we had at least some disposition to form, maintain, and revise these mental states in some of the ways that count as rational for mental states of those kinds. For example, Christine Korsgaard (1996a: 235) claims that conforming to the requirements of logic is constitutive of thinking, and conforming to the Kantian requirement of universalizability is constitutive of willing: as she puts it, “if I am going to think I must think in accordance with the principle of non-contradiction”, and “if I am going to will at all I must do so universally… The requirement of universality is in this way constitutive of willing.” The same basic idea has been developed in a different direction by philosophers who are sympathetic to decision theory: for example, David Lewis (1974) argued that we would not even be interpretable as having preferences at all unless our choices tended, by and large, to satisfy these basic axioms of decision theory (such as transitivity, monotonicity, independence, and so on).

It may be plausible to claim that the principles of rationality are constitutive of the mental states that they apply to in this way. If this claim also explained why we should conform to those principles of rationality, then we would have a solution to our problem – the problem of explaining why rationality matters – that is an instance of the second of the three responses to the problem that I listed above: that is, we would have an explanation of why rationality matters that appeals solely to internal facts about what is going on in the thinker’s mind.

However, it is not clear that this claim can explain why rationality matters. The claim that the principles of rationality are constitutive of the relevant mental states can be interpreted in several

---

9 I have defended this view of the constitutive role of principles of rationality myself; see Wedgwood (2007, chap. 7).

10 In a somewhat similar vein, Gauthier (1985) argues that unless one conforms to the axioms of decision theory, one will not count as a unified agent with a genuine will, as opposed to a bundle of disparate desires and needs.
ways. On one interpretation, it implies that we cannot have the relevant mental states unless we always conform to these principles of rationality. But on this interpretation, the claim implies that violating this requirement of rationality is impossible: if one does not conform to this requirement, then one is not in fact having an irrational mental state of the relevant kind at all. This is an intrinsically implausible implication, since irrationality seems to be a genuine phenomenon; it also seems in a way to trivialize this requirement of rationality if it is in fact a requirement that is utterly impossible for anyone to violate.

For this reason, it is more plausible to interpret this constitutive claim in a second way. According to this second interpretation, it is constitutive of having mental states of the relevant kind that one has a disposition, by and large, to conform to these requirements of rationality; it is not constitutively necessary that one must conform to these requirements in every single case. But this too seems not to provide the desired explanation. Suppose that it is true that it is constitutive of having preferences at all that one’s preferences must tend, by and large, to satisfy the coherence requirements that are implied by the axiom of decision theory. Is this really enough to explain why it is always irrational to have preferences that do not satisfy these requirements? Even if it is impossible for you to have preferences at all unless your preferences are disposed, by and large, to satisfy these requirements, it does not follow that they ought always to satisfy these requirements: perhaps it does not matter if you occasionally make choices that do not satisfy these requirements. Imagine trying to persuade a thinker to satisfy these requirements by pointing out that having some disposition to satisfy these requirements is necessary for having any preferences at all. The thinker might accept that having this disposition is necessary for having any preferences, but still wonder whether – given that it seems possible to fail to manifest this disposition in some cases – she should not sometimes resist this disposition.

So, the “constitutivist” approach also seems not to be sufficient to solve our problem. This is not to say that such constitutive claims cannot be part of the solution, but the solution requires more than just the truth of these constitutive claims. The failure of these two purely internal solutions to the problem suggests that we should consider an approach that invokes the idea that rationality has what could be thought of as an “external aim”.

4. The “external aim” of rationality

It is often been claimed that beliefs as such have an “aim” or “goal”: specifically, many philosophers have claimed that belief aim at the truth. A similar claim might also be made about other mental states and events as well: for example, perhaps choices and intentions aim at putting the chosen course of action in effect, in such a way that the agent thereby acts as she has most reason to do. Still, talking of an “aim” here is a metaphor. It is not immediately clear what the literal cash value of this metaphor is.

Strictly speaking, it is not just mental states that can be assessed as rational or irrational; we can also assess the rationality of mental events or processes, and of sets of mental states and mental events, as well. To keep things simple, however, in this discussion I shall just pretend that it is only mental states that we need to be concerned with here; this simplification will not affect any of the arguments that follow.

For our purposes, then, the important point is just that when a belief or some other such mental state achieves its “aim”, the mental state counts as correct – or in other words, as having got things right – and this external aim or standard of correctness that applies to a given type of mental state plays a crucial role in explaining the requirements of rationality that apply to this type of mental state. If the
requirements of rationality are explained by their relation to this external aim, then this relation to the external aim will presumably also explain why rationality matters, as I put it above – at least so long as this external aim is itself something that matters in an intelligible way.

In this way, the idea that these mental states have an “external aim” is a version of the third response to our problem (c): it implies that the explanation of why rationality matters refers to something external to the individual thinker’s mind. But this idea needs much further clarification and elaboration: in what way, exactly, are the internal requirements of rationality “explained by” their relations to the external aim of correctness?

There are in fact many different ways of conceiving of how the internal requirements of rationality are explained by their relations to an external aim of correctness. In the rest of this paper, I shall propose what now seems to me to be the most promising conception of this explanatory relationship. I shall try to make it clear that this conception of the relationship looks well placed to give a fully general explanation of the requirements of rationality (as opposed to an explanation of a small part of these requirements); but otherwise, I shall not be able to give a full defence of this conception here. My main goal is just to put forward this conception for further discussion.

This conception is based on the assumption that both irrationality and incorrectness come in degrees. Some beliefs and choices and processes of reasoning are more irrational than others; when we say simply that something (like a belief or choice or process of reasoning) is rational simpliciter, we can be taken to mean that it is no more irrational than any alternative that is available to the thinker at the time. Similarly, I shall assume that correctness comes in degrees: some beliefs and choices are more gravely or more badly incorrect than others; in that sense, some of these mental states have a worse – that is, greater – degree of incorrectness than others.

Now, according to the internalist assumptions that were introduced in Section 1 above, the rationality of your thinking at a given time depends purely on the mental states that are present in your mind at that time. Nonetheless, even though rationality depends purely on your mental states in this way, the relevant aspect of those mental states is what they are – to put it metaphorically again – “telling you” about the external world. The general connection between rationality and correctness is this: for your thinking to be as rational as possible, it must be the kind of thinking that has the best – that is, the lowest – possible degree of incorrectness according to what your mental states are “telling you” about the world.

Here we have another metaphor, in this talk of what your mental states are “telling you” about the world. What is the cash value of this metaphor? I propose that the content of this metaphor – what your mental states are “telling you” about the world – can be modelled by means of a space of worlds. This is because it is natural to spell out this notion of what your mental states are telling you in modal terms. There are some propositions that, according to what your mental states are telling you, must be the case: these propositions are true at all worlds in this space. There are also propositions that, according to what your mental states are telling you, might be the case: these propositions are true at some worlds in this space.

How can this space of worlds help us to understand what it is for one mental state $m_1$ to be less rational than a second mental state $m_2$? A simple way in which this might be the case is if $m_1$ is dominated by $m_2$ across this space of worlds: that is, if in every single world in this space, $m_2$ is no more incorrect than $m_1$, and in some worlds in this space $m_2$ is less incorrect than $m_1$. 


Two of the most famous attempts to explain why rationality requires probabilistic coherence – the “Dutch book” approach discussed above and the “accuracy dominance” approach pioneered by James M. Joyce (1998) – invoke this kind of dominance. That is, both of these approaches seek to explain requirements of rationality on the grounds that every set of mental states that violates these requirements is dominated by some alternative set of mental states that satisfies these requirements – whereas sets of mental states that satisfy the requirements are never dominated by any sets that violate them. (The main difference between these two approaches is that they assume very different measures of incorrectness: in effect, the Dutch book approach measures degrees of incorrectness in terms of net monetary losses, while the accuracy dominance approach measures degrees of correctness in terms of the closeness of the thinker’s degree of belief in the proposition to the proposition’s actual truth value.)

However, even if it is intuitively plausible that mental states that are dominated in this way are always irrational, it is not clear that this sort of dominance is what explains why they are irrational. The trouble is simply that this sort of explanation seems likely not to be sufficiently general. It is plausible that for some requirements of rationality, there are not only worlds where satisfying the requirement will result in your doing better in terms of correctness, but also some worlds where satisfying the requirement will result in your doing worse. For example, suppose that rationality requires you to be non-sceptical: that is, it requires you to take your sensory experience at face value – to respond to your having an experience as of a proposition p’s being the case (at least so long as you consider the question of whether p is the case, and no undermining or defeating evidence is present) by having a high level of confidence in p. Presumably, however, your mental states are not telling you that you could not possibly be a brain in a vat; so the space of worlds will include some worlds where you are a brain in a vat, and in those worlds, satisfying this requirement will lead you to do less well with respect to your beliefs’ degree of incorrectness than some ways of violating the requirement. So not every way of violating this requirement is dominated by some way of satisfying this requirement; this requirement cannot be explained by appealing to dominance in this way.

Still, it may be that the worlds in which satisfying this requirement leads you to do better, in terms of your beliefs’ degree of incorrectness, than every way of violating the requirement take up a larger proportion of this space of worlds than the worlds where violating the requirement leads you to do better in this way. This may be what explains why rationality requires us to take our sensory experiences at face value.

In general, to determine the degree of rationality of each mental state that the thinker could have, we could consider each of these mental states as partitioning this space of worlds into sub-regions or cells, such that for every one of these cells, the mental state in question have the same degree of incorrectness at every world in that cell. Then we could weight this degree of incorrectness by the proportion of the whole space of worlds that is taken up by the worlds in that cell; the average of these weighted degrees of correctness will be this mental state’s weighted average of incorrectness across this whole space of worlds. In general, then, I propose that one mental state is more rational than an alternative just in case the first state has better weighted average degree of incorrectness, across this whole space of worlds, than the second alternative state.

This conception of the relationship between correctness and rationality evidently requires this space of worlds to be a measurable space, in the sense that there are definite ratios between the different proportions of this space. So, for example, it could be that the worlds where one proposition p is true take up twice as large a proportion of this space as the worlds where another proposition q is true. In fact, as I shall propose in the next section, the relevant measure on this space is in effect a probability...
distribution over these worlds. In effect, then, the connection between rationality and correctness that I am proposing is simply this: \textit{rationality minimizes expected incorrectness} – where the relevant “expectation” is defined in terms of this probability measure on this space of worlds.\footnote{To make sense of a mental state’s “minimizing” expected incorrectness, we need to be able to make sense of the expected degree of incorrectness of mental states that one might possibly have, but does not actually have. There are some complications here, which I have explored elsewhere (Wedgwood 2013). But in general, I propose that to solve these complications, we should understand the degree of incorrectness of a (possible or actual) mental state as the degree of incorrectness that the mental state \textit{would have if} one had that mental state; and when I refer to the state’s expected degree of incorrectness, this should be understood as its \textit{conditionally} expected degree of incorrectness – conditionally on the assumption that one might express by uttering a present-tensed, first-person sentence ascribing that mental state to oneself.}

According to my proposal, then, all requirements of rationality – including both the requirements of rational belief and those of rational choice – are explained in the same way. First, for each kind of mental state there is some external “aim” – a standard of correctness that provides a way of measuring the degree of incorrectness that every mental state of that kind has at every possible world. Secondly, for every thinker at every time, there is a measurable space of worlds of the kind – where the measure on this space of worlds is in effect a probability measure. Finally, these two elements together determine an expected degree of incorrectness for every mental state; and according to my proposal, a mental state’s degree of irrationality can be identified with its expected degree of incorrectness.

I shall touch briefly on the first of these two elements – the external “aim” of each kind of mental state – towards the end of this discussion. The next two sections will be devoted to the second of these two elements – the probabilistically measurable space of worlds that represents what our mental states are “telling us” about the world.

5. \textbf{Why a probability distribution?}

According to the proposal that have just been made, the space of worlds relevant to the rationality of a thinker’s mental states and mental events at a given time is, as I put it, a “measurable space”. I shall now explain why it seems plausible that that the measure on this space of worlds must be a probability distribution. So far, I have only characterized this space of worlds in a rough intuitive way; if I give a slightly more rigorous characterization of this space of worlds, its probabilistic structure should become clear.

First, I shall assume that the propositions that can be built out of the concepts that the relevant thinker possesses form a \(\sigma\)-algebra on this space of worlds: that is, for every such proposition, there is a subset of the worlds in this space where that proposition is true, and the set of these propositions is closed under Boolean operations such as negation and disjunction and the like.\footnote{Since the number of possible worlds may be infinite, I shall not assume that there is a proposition corresponding to every set of worlds, or that every set of worlds takes up a well-defined proportion of the whole space. I assume only that there is such a set of worlds for every proposition, and each of these sets of worlds – the subsets that correspond to genuine propositions – takes up a well-defined proportion of the space.}

Secondly, I shall assume that there are definite \textit{ratios} between sub-regions of this space – for example, the worlds in one sub-region of the space might take up exactly \textit{twice as much} of the space as the worlds in another sub-region. In particular, there are also ratios between each sub-region of the space and the whole space of worlds itself. In this way, each subset of these worlds can be measured by numbers between 0 and 1, depending of the proportion of the whole space that is taken up by that subset: the empty subset takes up \textit{none} of this space, and so can be measured as 0, while the improper subset that takes up the \textit{whole} of this space can be measured as 1.
Finally, these ratios satisfy a fundamental condition of additivity: that is, the ratio between the region that is taken up by union of two disjoint sets of worlds $P_1$ and $P_2$ and the region that is taken up by a set of worlds $Q$ is the sum of the ratio between $P_1$ and $Q$ and the ratio between $P_2$ and $Q$. For example, if there are two disjoint subsets each of which takes up a quarter of the space, and so gets a measure of 0.25, their union will itself take up half of this space, and so get a measure of 0.5; and so on.

These assumptions are equivalent to the thesis that this measure on the space of worlds conforms to all the basic axioms of the probability calculus – and so itself counts as a probability distribution. I am proposing, then, that the space of possible worlds that corresponds to your mental states and mental events – the space that models or represents what your mental states and events are telling you about the world – itself essentially involves a probability measure on this space.

Why should this be true? Why is it that this space of possible worlds must have this sort of structure? Why could it not be a less structured heap of possible worlds instead? I have been describing this space of possible worlds as modelling or representing what your mental states are “telling you” about the world. So we might try to answer this question by showing that what your mental states are “telling you” about the world, in the relevant sense, cannot be adequately captured by an unstructured set of worlds, by only by a space of worlds that involves a probability distribution in this way.

However, we need to remember that in speaking of what your mental states “tell you” about the world, we have been talking metaphorically. (Literally speaking, “telling” someone something is a communicative act, and your mental states do not literally engage in communicative acts of any kind.) The underlying point here concerns the way in which what rationality requires of you depends on your mental states; and as I have insisted, the concept of rationality is itself a normative concept. Fundamentally, then, the point is simply that this is the sort of normative concept that the concept of rationality is: it stands for a certain distinctive virtue that mental states and processes of reasoning can exemplify; and it is part of the essential nature of this virtue that it assigns a measurable possibility space of this sort to the mental states and mental events and processes that are present in each thinker’s mind at each time.

It is, I suspect, a mistake to search for any deeper explanation of why rationality has this feature. Explanations have to start somewhere, and it is a mistake that philosophers sometimes fall into to attempt to explain something that is in fact so fundamental that it admits of no explanation at all. As Wittgenstein put it, eventually I will reach bedrock, and my spade is turned; the attempt to dig beneath explanatory bedrock is a futile endeavour.¹³

I suggest, then, that the fact that the virtue of rationality constructs a measurable possibility space of this sort out of your mental states is simply a basic feature of the essential nature of rationality. In this way, there is no deeper explanation of this fact about rationality; this fact forms part of the explanatory bedrock from which all normative explanations proceed. It seems intuitively plausible that rationality involves being guided by what is in some relevant sense possible, and what is not possible, given the mental states that one is having at the relevant time. The notion of probability, as I am using it here, is in effect just a notion of degrees of possibility; so it should also seem intuitively plausible that rationality involves being guided by these degrees of possibility as well.

¹³ See *Philosophical Investigations*, §217: “If I have exhausted the explanations, then I have reached bedrock, and my spade is turned.” Compare also *On Certainty*, §471: “It is so hard to find the beginning. Or better: It is hard to begin at the beginning, and not to try to go further back.”
Of course, we can still inquire what reason there is for us, as theorists, to believe that the virtue of rationality has this connection to probability. Even if this is an explanatorily fundamental feature of rationality, it should still be possible to offer a defence of the claim that rationality has this feature. A full defence of this claim would involve showing two things. First, it would involve showing that this claim about the fundamental structure of rationality provides the best explanation of a wide range of other phenomena that intuitively seem to call for explanation. Secondly, it would also involve showing that this claim about the structure of rationality (when properly understood) does not give rise to any special problems of their own – that is, it can answer all further objections that are likely to be raised against it. Unfortunately, I cannot offer a full defence of this sort here; the task of giving such a full defence of this claim must await another occasion.

At all events, there is certainly no problem in explaining why such a measurable space of possible worlds exists. There are clearly infinitely many measures of this sort that could be constructed on such sets of possible worlds. The question that clearly demands explanation is not why this measurable space of possible worlds exists, nor even why the concept of rationality picks out such a measurable space of possible worlds out of the thinker’s mental states, but rather why it picks out the particular measure that it does, rather than any of the infinitely many other measures that exist.

I am not suggesting that there is no answer to this question. On the contrary, it seems to me that there must be an explanation of why, out of all the probability distributions on these possible worlds that exist, it is this particular distribution that counts as relevant for determining how it is rational for the agent to think at the relevant time. In the next section, I shall try give a rough sketch of the sort of explanation that there could be of why it is precisely this probability distribution that is relevant in this way.

6. Why this probability distribution?

As I have put it, metaphorically, the relevant space of worlds must reflect what the relevant thinker’s mental states and events are telling her about the world. Strictly, the thinker’s mental states may not determine a unique probability distribution over this space of worlds. What the thinker’s mental states are telling the thinker need not be quite as determinate as a unique probability distribution. In such cases, there will simply be a large set of probability distributions each of which reflects what the thinker’s mental states are telling her equally well. The most plausible way to revise my proposal to accommodate such cases would be by saying that one mental state is less rational than another alternative state if and only if the first state has a worse expected degree of incorrectness according to every probability distribution in this set.

In what follows, however, I shall ignore this complication; I shall write as though the thinker’s mental states determine a unique space of worlds and a unique probability distribution over that space. The question that I shall explore in this section is how exactly this space and this probability distribution are determined by the thinker’s mental states.

The intuitive idea is that there are some connections between these mental states and the truth (including the truth about the external world) that are essential to these mental states – that is, these connections to the truth are somehow built into the constitutive nature of those mental states. This notion of the constitutive nature of a mental state needs to be understood in a certain way. Consider the concepts that you express by the words ‘Hesperus’ and ‘Phosphorus’. Even if it is a necessary truth that these two concepts refer to the same object, that necessary truth is grounded, at least in part, in the facts of astronomy and in the necessity of identity, not simply in the nature of these concepts as
constituents of your mental states themselves. In that way, it is not part of the constitutive nature of these concepts themselves that they refer to the same object. This is what makes it possible that your current mental states might not tell you that Hesperus = Phosphorus.

To accommodate this feature of what your mental states are telling you (and what they are not telling you), the relevant space of worlds must contain some worlds where Hesperus ≠ Phosphorus. Thus, some of these worlds will be metaphysically impossible. The easiest way to picture this is to suppose that each world is simply a set of Fregean propositions or Gedanken (that is, propositions that are built up, not out of objects, properties and relations themselves, but out of concepts or modes of presentations of those objects, properties and relations); to say that a proposition p is true at a world w is simply to say that the world w is a set of propositions that includes that proposition p.

On this way of picturing things, then, some of these worlds include the Fregean proposition that Hesperus ≠ Phosphorus (which is distinct from the Fregean proposition that Hesperus ≠ Hesperus). The Fregean proposition that Hesperus ≠ Phosphorus is metaphysically impossible, but it is a real proposition nonetheless; and since your mental states may not be telling you that it is false, there should be some worlds in the space that where this proposition is true.

However, if p is a conceptual truth, built up out of concepts that you possess, then your mental states are in effect telling you that p is true. The concepts that are involved in your mental states in some sense “guarantee” the truth of these conceptual truths. (For example, your mental states are in the relevant sense “telling you” such things as the following: that all bachelors are unmarried; that if someone knows a proposition p, then p is true; and so on.) Thus, these conceptual truths must hold at every single world in the space.

Similarly, if p and q are conceptually incompatible with each other, your mental states are in effect telling you that p and q are not both true (the nature of the concepts guarantees that they cannot both be true); and so p will not be true at any world at which q is true, and vice versa. So every world in this space is internally consistent: there are no worlds where any two conceptually incompatible propositions are both true. Finally, if p conceptually entails q, then your mental states are conditionally telling you that q is true, conditionally on the assumption that p is true; so q must be true at every world in this space where p is true. Thus, every world in this space is closed under conceptually valid entailment: for example, every world where Eric is a bachelor is a world where Eric is unmarried; and so on.

In this way, the worlds in this space must all be epistemically possible worlds: they are world that respect all conceptual truths in this way. Given that all logical truths are conceptual truths, it follows that every world must be logically possible; the probability measure on this space of worlds must assign probability 1 to every logical truth and 0 to every logical falsehood, and it must always assign the same probability to any two logically equivalent propositions.

Besides these conceptual truths, another thing that your current mental states are “telling you” about the actual world is that you are currently having those very mental states. So, every true proposition (built up out of concepts that you possess) about the mental states that are currently present in your mind must also be true at every epistemically possible world in this space. So, for example, if you are currently in pain, and you have the concepts to think the proposition that you are now in pain, then in every world in this space, the proposition that you are now in pain is true. This is not to say that every true proposition about your mental states is true in every one of these worlds: propositions about the mental states that you had in the past, or will have in the future, will not typically be true in every
such world; and if there are truths about your current mental states if they cannot be captured in the concepts that you currently possess. But the conceptually accessible truths about your current mental states will hold throughout this space of worlds.

So far, what I have said about this space of epistemically possible worlds is reminiscent of the space of worlds that is postulated in the epistemological work of David Lewis (1999, Essay 25). However, I am assuming that this space of worlds involves a probability distribution over the space, besides a mere set of worlds, and in this way has more structure than the kind of space that Lewis discusses. Somehow this probability distribution is determined by what is present in the mind of the thinker at the relevant time. But how could the thinker’s mental states determine this probability distribution?

One thing that our mental states seem not to be telling us is that every possible world is equally probable. For this reason, we should certainly not assume that the probability measure on the space of worlds has to be a uniform measure. Some possible worlds may be more probable than others. (In fact, this will have to be the case if there are infinitely many possible worlds.) Some possible worlds may even be maximally improbable: that is, even though they are epistemically possible worlds, each of these worlds may be thought of as taking up no more than an extensionless point in this space.14 Similarly, certain sets of worlds may also be thought of as taking up such an extensionless point in this space; the propositions that are true only at such sets of worlds are epistemically possible propositions, even though their probability is 0. (For example, propositions about the precise real value taken by an empirical quality may be instances of such 0-probability possible propositions.)

In fact, many familiar views about what rationality requires of us can in effect be reinterpreted as views about the structure of this space of possible worlds. For example, consider the so-called “Principal Principle”. According to this principle, your conditional credence in a proposition A, given the supposition that the chance assigned to A by the true theory of the world T and the true history of the world H up to a certain point in time t is x, should itself be x, so long as you have no “inadmissible information” regarding A – that is, no information that does not follow from the true theory T and history H (such as information about which chancy events happen at times later than t).15

The Principal Principle can be encoded as a condition on this space of worlds. Consider the sub-region of the space where it is true that a certain event (for example, a certain outcome of a coin toss) has a certain objective chance; and suppose that no “inadmissible information” is true at all worlds in the sub-region. Then the proportion of this sub-region taken up by worlds where that event occurs will correspond to that objective chance. This will ensure that this probability measure obeys the Principal Principle. Is this a reasonable constraint to impose on this probability measure? It will be reasonable if it is plausible that whenever the hypothesis that the objective chance of the event in question is x is epistemically possible, and you have no inadmissible information regarding that event, then one of the things that your mental states are conditionally telling you, conditionally on the assumption that this hypothesis is true, is that the degree to which this event is possible corresponds precisely to x.

In general, many of the great questions of epistemology can be reconceived as questions about the properties of this space of epistemically possible worlds. I suggested above that the explanation of why the conceptual truths built into the nature of the concepts that you possess must hold throughout this space of possible worlds has something to do with the way in which the nature of these concepts somehow “guarantees” the truth of these conceptual truths. It may be possible to extend this kind of explanation to account for other features of this space of worlds as well – although the explanation

14 For an illuminating discussion of this point, see Hájek (2012).
15 For a discussion of how exactly to understand the Principal Principle, see Meacham (2010).
might have to focus on other mental types and properties besides concepts, and it might have to invoke a weaker connection to the truth than that of “guaranteeing” the truth of the relevant propositions.

For example, it may be a feature of this space of possible worlds that if in every world in this space you have the kinds of experiences that I am currently having, then in most of this space – though not quite all of this space – you really are sitting on a chair typing at a computer keyboard (rather than dreaming or being deceived by an evil demon or the like). It may be possible to provide an explanation of the kind that I have described of why the space of possible worlds has this feature. This explanation would probably have to appeal not only to the nature of the relevant concepts, but also to the nature of relevant types of sensory experiences; and the essential connection that these experiences and concepts have to the truth seems to be weaker than that of guaranteeing that these propositions are true. Nonetheless, it may be that this connection to the truth is enough to explain why the relevant space of possible worlds must have this feature.  

The precise details of this explanation do not matter for our present purposes. Since it seems rational to have a high degree of confidence in the contents of one’s sensory experiences (at least in the absence of special defeating factors), there is presumably some explanation that can be given of why it is rational. Whatever this explanation is, it should be capable of being adapted to give an explanation of why this space of possible worlds must have this feature. So the precise features of this space of possible worlds should be no less explicable than these highly plausible features of rationality. In this way, then, many of the great questions of epistemology can be reinterpreted as questions about the probability measure on this space of worlds. It is the answers to these epistemological questions that will determine what this probability distribution is like.

Some readers might be inclined to take my talk of what the thinker’s mental states are “telling her” to point towards a highly subjective conception of this probability distribution. According to such a subjective conception, the probability distribution simply corresponds to the degrees of belief that the believer actually has – or at least corresponds as closely as possible, given that the believer’s actual degrees of belief may not be perfectly probabilistically coherent.  

In fact, however, I have already suggested that there are other constraints on this space besides bare probabilistic coherence itself: for example, the space may have to encode the Principal Principle; or it may encode the non-sceptical principle according to which we are rationally required to take our sensory experiences at face value (so long as no defeating evidence is present). But the thinker’s beliefs might violate the Principal Principle, or the thinker might fail to take her sensory experiences at face value. So this space of worlds does not simply correspond to the beliefs that the thinker actually has; those beliefs might fail to reflect exactly what the thinker’s mental states are really telling her about the world.

Other readers may have noticed that I have said that the probability distribution on this space of worlds is determined by the thinker’s mental states and mental events, and I have also suggested that every true proposition about the thinker’s current mental states built up out of concepts that the thinker possesses must be true in every world in this space – thus ensuring that every such proposition has probability 1. This might suggest to these readers that the probability distribution is a kind of “evidential probability”.  

If these propositions about the thinker’s mental states are equated with the thinker’s “evidence”, then perhaps this probability distribution could be identified with the result of conditionalizing a special privileged Ur-probability function on the thinker’s “evidence”?  

---

16 I have tried to develop this sort of approach in more detail elsewhere; see Wedgwood (2011).
17 For a classic statement of this subjective Bayesian position, see Jeffrey (2004).
18 For a development of this idea of “evidential probability”, see Williamson (2000, chap. 9).
In fact, however, I need not commit myself to any such evidential probability. The reason for this is that there is more to what is going on in your mind than can captured by the true propositions, involving concepts that you possess, about your current mental states. First, not all truths about what is going on in your mind can be captured by the concepts that you possess. For example, your sensory experiences may have non-conceptual contents; and these non-conceptual contents may be part of what determines what your mental states are telling you, even if your mental states are not telling you that you have are having sensory experiences with precisely those non-conceptual contents. Secondly, part of what is going on in your mind is a transition to your current mental states from your immediately preceding states, and your current mental states need not be telling you about your past mental states. The nature of this transition, from your past mental states to your current states, may also be part of what determines what your mental states are telling you, even if your states are not telling about the precise nature of this transition.

Thus, there could be aspects of what is going on in your mind, which are not captured by these true propositions about your current states to which this probability distribution assigns probability 1, that play a role in determining what the probability distribution itself is like. If so, then it seems that we also cannot identify this probability distribution with any kind of “evidential probability”.

7. Epistemic utility theory?

At all events, my central proposal is that rationality optimizes – that is, minimizes – expected incorrectness, where the “expectation” is defined in terms of this probability distribution on the relevant space of epistemically possible worlds. According to this proposal, the notion of maximizing the expectation of some value is part of the very essence of rationality.

This sort of approach is often conceived as reducing all of rationality to practical rationality or decision theory – and so reconceiving epistemology as a kind of “cognitive decision theory” or “epistemic utility theory”. It seems to me, however, that this is not the best way to think of this approach.

The notion of something’s expected value is defined in terms of two functions – a probability function and a value function. Classical decision theory, in the tradition of Leonard Savage (1972) and Richard Jeffrey (1981), has conceived of both the probability function and the value function subjectively, as psychological measures of an agent’s beliefs and desires respectively. According to classical decision theory, the relevant probability function is a credence function (a measure of the agent’s partial beliefs or degrees of confidence) and the relevant value function is a utility function (a measure of the agent’s subjective preferences).

The general conception of rationality that I am proposing here interprets these two functions in a significantly different way. Specifically, this conception does not interpret these functions as psychological measures of the beliefs and desires that the agent actually has. Instead, it interprets them as measures of essentially normative phenomena – namely, of what the agent’s mental states are in a normative sense “telling” her about the world, and of how far various possible mental states and events fall short of being perfectly correct or as they ought to be. In this way, this conception interprets both the probability function and the value function as in a sense more objective than the functions that are invoked by the decision theories of Savage and Jeffrey.

---

19 For example, the term ‘epistemic decision theory’ is used by Greaves (2013), while Pettigrew (2012) uses the term ‘epistemic utility theory’.
As I have explained, the value function is a measure of degrees of incorrectness; and, I propose, this standard of correctness is fixed by the fundamental norms that apply to each kind of mental state—such as belief, or choice, or the like—and not directly by the preferences that the agent happens to have. Similarly, the probability function is a probability distribution on the relevant space of worlds—which is also, according to this conception, an irreducibly normative phenomenon. Admittedly, this probability distribution over this space of worlds is determined by the thinker’s mental states somehow; this point follows from the internalist assumptions that were articulated in Section 1. However, as I explained in the previous section, this probability distribution need not exactly correspond to the beliefs that the thinker actually has. If the thinker’s beliefs fail to cohere in the rationally required manner with each other or with the thinker’s sensory experiences, memories, and other mental states, then the thinker’s beliefs are at least to some extent irrational. In that case, what the thinker’s mental states are really “telling her” is significantly different from what is explicitly encoded in her beliefs about the world.

Rejecting this subjectivist understanding of the relevant value functions is, in my view, indispensable for making it plausible that the notion of an expected value is fundamental to epistemology as well as to decision theory—since it does not seem at all plausible that the agent’s subjective desires or preferences can make such a radical difference to what it is rational for the agent to believe. Moreover, in rejecting the subjective notion of utility in favour of the more objective normative notion of the degrees of correctness or incorrectness that a choice might have, I am deliberately aiming to bring decision theory closer to epistemology. The goal behind this proposal, then, is not to reduce epistemology to decision theory, but simply to reveal the common structure that is present in both rational belief and rational choice.

On this picture, then, the principle that rationality minimizes expected incorrectness is the fundamental principle that gives the essential character of the virtue of rationality: all the more specific principles of rationality applying to each type of mental state are explained by this general conception, and by features of the relevant space of possible worlds, together with the principle that defines what it is for mental states of that type to be correct. For example, the specific principles of rational belief are explained by this general conception of rationality together with the principle that defines what it is for beliefs to be correct; and the specific principles of rational choice are explained by this general conception together with the principle that defines what it is for choices to be correct.

On this picture, then, the notion of “correctness”—the notion of the external “aim” of each type of mental state—is of paramount importance for understanding the requirements of rationality. I cannot give a full account of this notion of “correctness” here, but in the last section, I shall briefly comment on what the central proposal that I have made here can tells us about this external standard of evaluation for mental states.

8. What this tells us about the external aim of rationality

As I shall argue in this final section, the central proposal that I have made in this paper in fact gives us a crucial clue about what the external aim of each sort of mental state must be like. The reason for this is that this proposal seems to lead directly to the idea of a certain special sort of requirement: I shall call these requirements “aim-reflecting requirements” (although as we shall see, they could perhaps equally well be called “anti-akrasia requirements”).
Suppose that it is rational for you to be *certain* that a certain attitude \( A_1 \) is more incorrect – further away from the relevant “aim” – than another attitude \( A_2 \). Then presumably it must be true throughout this space of epistemically possible worlds that \( A_1 \) is more incorrect than \( A_2 \). So, whatever the probability measure on this space of worlds may be, the expected incorrectness of \( A_1 \) must be greater than that of \( A_2 \). So, given the proposal of Section 4, this attitude \( A_1 \) cannot minimize expected incorrectness; \( A_1 \) must therefore be irrational. If every alternative to \( A_2 \) is irrational in this way, then \( A_2 \) itself must be rationally required.

In fact, if we can make sense of having conditional versions of these attitudes \( A_1 \) and \( A_2 \) (as we can make sense of conditional beliefs, and arguably also conditional choices or intentions), then it seems that we are rationally required not to have the conditional version of \( A_1 \) conditional on the assumption that \( A_1 \) is more incorrect than \( A_2 \); and we are rationally required to have the conditional version of \( A_2 \) conditionally on the assumption that every alternative to \( A_2 \) is more incorrect than \( A_2 \) is itself.

In the case of belief, this implies, in effect, that given the assumption that the most correct belief for one to have about \( p \) is to believe \( p \), one is rationally committed to having a conditional belief in \( p \). That is, one is rationally required to accept the *inference* from the assumption that one could express by saying ‘The most correct belief for me to have about \( p \) is to believe \( p \)’ to the conclusion \( p \) itself. Thus, it seems that one is committed to accepting that the proposition that one could express by saying ‘The most correct belief for me to have about \( p \) is to believe \( p \)’ implies the proposition \( p \) itself. In this way, it seems, we seem to be led to the conclusion that for a belief to be maximally correct, the proposition believed must be true.

A similar point applies in the case of choice and intention. Given the assumption that the most correct choice for one to make between a set of alternatives is to choose option \( A \), one is rationally committed to making a conditional choice for \( A \) (conditionally on this assumption’s being true). This requirement should be one of the *basic* requirements of rationality. There should be something more or less *obviously* irrational about violating these requirements; sane and intelligent agents must all have at least a tendency to avoid violating such requirements.

The only way to interpret this requirement as a basic requirement of this sort is to interpret it as a requirement that rules out a kind of *akrasia* – that is, as a requirement that rules out choosing one of the alternatives to \( A \), when one is certain that one has most reason, all things considered, to do \( A \), and that one’s doing \( A \) depends on one’s choosing to do it. In short, the proposition that the most correct choice for one to make is to choose \( A \) implies that – at least if doing \( A \) depends on one’s choosing to do it – then \( A \) is what one has most reason to do. In this way, we are led to the conclusion that a choice will be maximally correct whenever (i) the option chosen is what one has most reason to do, and (ii) one’s performing that option depends on one’s making the choice.

This idea of the external aim of mental states like belief and choice clearly requires much more extensive exploration. In this section, I have simply tried to show that the central proposal that I have made in this paper gives some crucial clues about what the external aim of each of these kinds of mental states is. According to this proposal, this external aim defines an externalist standard of correctness and incorrectness for each type of mental state. Rationality differs from correctness in being an internalist standard, rather than an externalist standard. However, rationality and correctness have a fundamental probabilistic connection to each other: the mental states and events present in the thinker’s mind at the relevant time determine a certain probability distribution over a space of

\[ \text{Footnote 11} \]

20 In fact, this formulation is not quite exact. It needs to be amended to deal with the complications that were broached in footnote 11 above.
epistemically possible worlds; and to think in a rational way is to think in a way minimizes the expectation, according to this probability distribution, of one’s thinking’s degree of incorrectness.

References