Philosophy and Phenomenological Research
Vol. LXV, No. 2, September 2002

Internalism Explained*

RALPH WEDGWOOD
Merton College, Oxford

1 Explaining Internalism

The word 'rational' is used in many ways. But when the word is used in the way that is most common among philosophers, the following intuition seems compelling.

Consider two possible worlds, w_1 and w_2. In both worlds, you have exactly the same experiences, apparent memories, and intuitions, and in both worlds you go through exactly the same processes of reasoning, and form exactly the same beliefs. In this case, it seems, exactly the same beliefs are rational in both worlds, and exactly the same beliefs are irrational in both worlds. Now suppose that in w_1 you are bedevilled by an evil demon who ensures that many of your experiences are misleading, with the result that many of the beliefs that you hold in w_1 are false. In w_2, on the other hand, almost all your experiences are veridical, with the result that almost all the beliefs that you hold in w_2 are true. Intuitively, this makes no difference at all. Exactly the same beliefs are rational and irrational in both worlds.[1]

This intuition seems to support an "internalist" conception of rational belief. According to this conception, the rationality of a belief supervenes purely on "internal facts" about the thinker's mental states—in this example, on facts that hold in *both* these two possible worlds w_1 and w_2, not on facts about the external world that vary between w_1 and w_2.

Moreover, this seems to be a completely *general* feature of rationality: it is not just rational beliefs that have this feature; the same feature seems

---

* **Editor's Note:** This essay won the Young Epistemologist Prize, sponsored by Rutgers and PPR, in 2001.

1 Several epistemologists have proposed theories that are incompatible with this intuition (at least assuming that a belief is "justified" if and only if it is rational). See, e.g., Timothy Williamson, *Knowledge and its Limits* (Oxford: Clarendon Press, 2000), chapter 9, and Alvin Goldman, "What is Justified Belief?", in George Pappas, ed., *Justification and Knowledge* (Dordrecht: Reidel, 1979). Even Goldman, however, seems to have felt the pull of this intuition, when he suggested that the rules that it is rational for us to follow, in forming and revising our beliefs, are the rules that are most reliable in "normal worlds"—that is, worlds in which our general beliefs are true; see his *Epistemology and Cognition* (Cambridge, Massachusetts: Harvard University Press, 1986), pp. 107–09.

present in rational choices or decisions as well. When we assess a choice or decision as rational or irrational, we are assessing it on the basis of its relation to the agent's beliefs, desires, and other such mental states—not on the basis of its relation to facts about the external world that could vary while those mental states remained unchanged.[2] This also seems to be a *special* feature of rationality, in contrast to other ways of evaluating beliefs and decisions. All the other ways of evaluating beliefs and decisions—for example, as "correct" or "incorrect", "advisable" or "inadvisable", and so on—are externalist evaluations. What is distinctive of "rationality" (at least as the term is most commonly used by philosophers) is that it is an internalist evaluation.

Thus, internalism with respect to rationality seems to have considerable intuitive support. However, this intuitive support would be undermined if we cannot give an adequate *explanation* of internalism. Specifically, we must explain exactly *which* facts about a thinker count as these "internal facts" upon which the rationality of a belief or decision supervenes. We must also explain *why* rationality should supervene on internal facts in this way. In §2 of this paper, I shall argue that the standard version of internalism fails to provide satisfactory explanations here. In §3, I shall propose an alternative conception of rationality; and in §§4–5, I shall argue that this alternative conception provides a better explanation of what exactly these "internal facts" are, and of why it is that the rationality of beliefs or decisions supervenes on them.

As I have remarked, if internalism is true, then it applies just as much to rational decision as to rational belief. However, the question of whether or not internalism is true has chiefly been discussed by epistemologists, rather than by theorists of practical reason. For this reason, I shall focus exclusively on rational belief here, and ignore rational decision altogether. But this is a purely stylistic choice on my part. All my arguments would work just as well if they focused on rational decision instead of rational belief.

2 The Standard Version of Internalism

According to internalism, the rationality or irrationality of a belief is determined purely by "internal facts" about the thinker. As James Pryor puts it, the standard version of internalism defines these "internal facts" as "*facts to*

---

[2] This formulation is designed to be compatible with externalism with respect to *mental content*—that is, the view that the content of a thinker's beliefs and desires does not supervene on narrow, intrinsic properties of the thinker alone, but may also depend on the thinker's relations to her environment. For the classic argument in favour of externalism with respect to mental content, see Hilary Putnam, "The Meaning of 'Meaning'", reprinted in his *Mind, Language, and Reality* (Cambridge: Cambridge University Press, 1975).

which one has a special kind of access."[3] Specifically, one has this "special kind of access" to a fact just in case one is in a position to know that fact "by reflection alone". In this context, 'reflection' means "*a priori* reasoning, introspective awareness of one's own mental states and one's memory of knowledge acquired in these ways." In short, according to this standard version of internalism, whether or not it is rational for one to believe a proposition "supervenes on facts that one is in a position to know about by reflection alone".

Many proponents of this standard version of internalism attempt to explain *why* internalism should be true by claiming that to say that a belief is "rational" is just to say that in holding that belief, the thinker is proceeding in a "cognitively blameless" fashion. Since it seems that one cannot fairly be blamed for not responding to a fact that one was not in a position to know, this point is held by some philosophers to explain why an internalist conception of rationality must be true.[4]

This standard version of internalism is open to grave objections. First, the claim that rationality is simply a matter of mere "cognitive blamelessness" seems false. There are at least two ways in which an act can be "blameless"—either because the act is *justified*, or because it is *excusable*. For example, if I kill you because it was the only way for me to defend myself against your attempt to murder me, my act may be justified; if I kill you because I have gone insane, my act may be, not justified, but excusable.[5] Roughly, to say that an act is justified is to say that the act should be approved of; to say that an act is excusable is to say that, although the act should not be approved of, the agent should not be blamed for having done it. Clearly, the notion of rational or justified belief is much closer to the notion of a justified act than to the more general notion of a blameless act. Thus, not all "cognitively blameless" beliefs are rational or justified: a belief can be blameless merely because it is excusable, even if it is not rational or justified in any way.[6]

Moreover, as Alvin Goldman has recently argued,[7] it is doubtful whether the claim that rationality is just a matter of "cognitive blamelessness" does

---

3 Pryor, "Highlights of Recent Epistemology", *British Journal for the Philosophy of Science* 52 (2001), 95–124, pp. 103–4.

4 For an illuminating discussion of the possibility of explaining internalism in this way, see William Alston, *Epistemic Justification* (Ithaca, New York: Cornell University Press, 1989), Essay 8.

5 The distinction between justification and excuse plays an important (albeit contested) role in English and American criminal law; see, e.g., Michael L. Corrado, ed., *Justification and Excuse in the Criminal Law* (New York: Garland, 1994).

6 Variants of this point have been made by Pryor, "Highlights of Recent Epistemology", pp. 114–18, and also by Alvin Plantinga, *Warrant: The Current Debate* (New York: Oxford University Press, 1993), p. 39.

7 Goldman, "Internalism Exposed", *Journal of Philosophy* 96 (1999), 271–93.

explain this version of internalism. Even if one cannot fairly be blamed for not responding to a fact that one was not in a position to know, it is much less plausible to claim that one can never fairly be blamed for not responding to a fact that one was not in a position to know "by reflection alone".

Anyway, once "internal facts" are defined in this way, it is also doubtful whether internalism is true. As Timothy Williamson has recently argued, there seems not to be any domain of non-trivial facts such that it is an *essential* feature of facts within that domain that one is in a position to know those facts by reflection alone.[8] Thus, the following worlds w_1 and w_2 both seem possible. In both w_1 and w_2, you believe p on the basis of certain reasons, but in w_1 you are in a position to know by reflection alone that you believe p on the basis of those reasons, while in w_2 you are not in a position to know this; otherwise, you are in just the same mental states in both w_1 and w_2. So, according to the standard version of internalism, the fact that you believe p on the basis of these reasons may be part of what makes the belief rational in w_1, but it cannot be part of what makes the belief rational in w_2. Hence, this version of internalism must say that it could be the case that this belief is rational in w_1 but not rational in w_2. But then the fact that you are in a position to know about the basis for your belief in w_1 is *itself* one of the facts on which the rationality of the belief supervenes. The fact that you are in a position to know these facts is not merely a *precondition* of these facts' determining whether the belief is rational. It is *itself* one of the facts that determine whether the belief is rational. So, according to this standard version of internalism, you must also be in a position to know *that* fact by reflection alone. Thus, if there is any set of facts that determines whether a belief is rational, then for every fact F that belongs to that set of facts, that set must include, not just F, but also the fact that one is in a position to know F by reflection alone, the fact that one is in a position to know by reflection alone that one is in a position to know F by reflection alone, and so on, *ad infinitum*. This makes it doubtful whether there is any set of facts that determines whether any belief is rational at all.

For these reasons then, both components of the standard version of internalism seem dubious. First, it seems that rationality must involve more than mere "cognitive blamelessness". Second, it seems not to be necessary that the "internal facts" that determine whether or not a belief is rational should all be facts that one is "in a position to know by reflection alone"; these "internal facts" must be defined in some other way.

3 Following Basic Rules

In this section, I shall propose a certain conception of what it is for a *belief revision* to be rational. I shall use the phrase 'belief revision' broadly, so that

[8] Williamson, *Knowledge and its Limits*, chapter 4.

it includes not only forming a new belief, but also reaffirming or abandoning an old belief. Even if one is not currently forming or reaffirming a belief, one may still hold that belief as a *background belief*—that is, as a standing mental state stored in propositional memory. However, to simplify the discussion, I shall just focus on the question of what it is for belief revisions to be rational. I shall ignore the question of what it is for background beliefs to be rational here.[9]

It is often claimed that whenever we revise our beliefs, we are thereby pursuing some *aim*. For example, some philosophers claim that, whenever one forms or reaffirms or abandons one's belief in a proposition p, one is thereby pursuing the aim of believing p if and only if p is true. It is not completely clear how to interpret such claims. Are these claims literal or metaphorical? If they are metaphorical, how exactly is the metaphor to be interpreted? But let us assume that there is some reasonable interpretation on which it is true to claim that whenever one revises one's beliefs, one is thereby pursuing some aim. It does not matter for my purposes exactly *what* aim one is thereby pursuing. To fix ideas, however, let us assume that whenever one revises one's beliefs in a proposition p, one is pursuing the aim of believing p if and only if p is true. How is one to pursue this aim?

Presumably, to pursue this aim, one must revise one's beliefs in certain ways when one is in certain conditions, and revise one's beliefs in other ways when in other conditions. We may imagine a set of *rules*, such that each of these rules permits one to revise one's beliefs in a certain way whenever one is in a certain related condition. For example, one such rule might permit[10] one to come to believe p whenever one has an experience as of p's being the case (and one has no special reason to distrust one's experiences in the circumstances). One would "conform" to such a rule just in case one revises one's beliefs in a certain way at the same time as being in a certain condition, and the rule in question permits one to revise one's beliefs in that way when

---

9 The notions of a rational background belief and of a rational belief revision seem to be connected in the following way: so long as it is rational for a thinker to hold a certain background belief, then no belief revision on the part of the thinker will be irrational merely because it is based on that background belief. Given this connection between rational belief revisions and rational background beliefs, internalism about belief revisions will entail internalism about background beliefs too. For an illuminating discussion of when such background beliefs are rational, see Alan Millar, *Reasons and Experience* (Oxford: Clarendon Press, 1991), chapter 6.

10 Isn't one *rationally required* (not merely *permitted*) to form this belief when in this condition? Yes, but we do not need to add that this rule requires (as opposed to merely permits) forming this belief in this condition. We need to add that there is no rule that permits one to disbelieve p, or to suspend judgment about whether p is the case, when in this condition. As I shall later propose, a belief revision is rational only if it conforms to one of these rules. This is why it is not rational for one to do anything other than believe p when one has an experience as of p's being the case (and no special reason to distrust one's experiences in the circumstances).

in that condition. Then perhaps there is a certain set of such rules that it "makes sense" for one to conform to, in order to pursue this aim. As I shall understand this phrase, to say that it "makes sense" to conform to these rules, in order to pursue this aim, is both to state a certain fact about these rules, and also to *recommend* conforming to these rules as a means to that aim. Admittedly, it is still unclear exactly what fact is stated here, and what sort of recommendation this is. We shall return to that question later. First, however, we need to be clearer about what is meant by describing one's conforming to these rules as a *means* to this aim—that is, as something that one *does, in order to* pursue this aim.

As I have described these rules, it is perfectly possible to conform to these rules by pure fluke. But if it is purely a fluke that one conforms to a rule, it will hardly be appropriate to say, even metaphorically, that conforming to the rules is something that one does, in order to pursue this aim. This description will be appropriate only if one not only conforms to the rule, but also *follows*, or is *guided by*, the rule. For example, consider the rule that permits one to come to believe p whenever one has an experience as of p's being the case (and no special reason to distrust one's experience in the circumstances). One would certainly not count as "following" this rule if it were simply a fluke that one comes to believe p at the same time as one has an experience as of p's being the case. At the very least, it must also be the case that one comes to believe p precisely *because* one has an experience as of p's being the case, and because this belief has the same content—p—as one's experience.[11] In general, if one follows a rule that permits one to revise one's beliefs in a certain way whenever one is in a certain related condition, then one revises one's beliefs in the relevant way *in response to* the fact that one is in the relevant condition.

It might seem obvious which rules it "makes sense" for one to conform to, in order to pursue the aim of believing the proposition p if and only if p is true. One should just conform to the "truth rule"—the rule that permits one to believe p if and only if p is true. But even if it is sometimes possible for one to follow this "truth rule", it may be that one follows this rule *by means of* following *other* rules. For example, one way to follow this "truth rule" might be by means of following the rule that permits one to believe p whenever one has an experience as of p's being the case (and no special reason to distrust one's experience in the circumstances).

So, at least sometimes, one follows some rules by means of following other rules. But if one follows any rules at all, then one must follow some

[11] This is still only a necessary condition, not a sufficient condition, for following this rule. I shall later add a further necessary condition in §5 below. I shall not attempt to establish here whether these conditions are jointly sufficient as well as necessary; the task of giving such necessary and sufficient conditions for "following a rule" is too large a question to be addressed here.

rules *directly*—not by means of following any other rules. Following a rule "directly", in this sense, is analogous to performing a *basic action*. A basic action is an action that one performs, but not by means of performing any other action.[12] If a thinker is able to follow a certain rule directly, in this sense, at a given time, then I shall say that the rule in question is a "basic rule" for that thinker at that time.

Some of the rules that are in this sense "basic rules" for a thinker at a given time will be rules that it "makes sense" for the thinker to conform to at that time, in order to pursue the aim of believing the proposition p if and only if p is true (or whatever exactly the relevant aim of revising one's belief in p may be). It is these rules, I propose, that are the *rules of rational belief revision* for the thinker at that time with respect to that proposition. If the thinker revises her belief in p at that time, then that belief revision is rational just in case it results from her directly following some of these basic rules that it "makes sense" for her to conform to.[13]

In this paper, I shall argue that this proposal provides an explanation for internalism. But in fact, this proposal may also explain another common intuition about rationality. Intuitively, the fact that one's belief is true, or counts as knowledge, is not something that lies within one's direct control; in that sense, it is partly a matter of good luck or good fortune. On the other hand, the fact that one's belief is rational is something that lies within one's direct control; it cannot be a matter of mere good luck or good fortune. One way of explaining this distinction, between what "lies within one's direct control" and what is "partly a matter of luck", is based on the notion of "basic actions". Clearly, for every basic action, it may be a matter of luck that one has the ability to perform that basic action: it may be a matter of luck, for example, that one is not paralysed, or insane, or dead. But *if* one has the ability to perform a certain basic action, then whether or not one performs that basic action lies within one's direct control. On the other hand, even if one is able to bring about a certain further result by means of performing that basic action, whether or not one actually brings about that result may be partly a matter of luck.[14] Similarly, if one has the capacity to follow a basic

---

12 See especially Arthur Danto, "Basic Actions", in Alan White, ed., *The Philosophy of Action* (Oxford: Oxford University Press, 1968), pp. 43–58.

13 For some more of the details of this conception of rationality, and in particular, for an account of how internalism can be reconciled with the idea that belief revisions have an external "aim" (such as truth and the avoidance of error), see my "The A Priori Rules of Rationality", *Philosophy and Phenomenological Research* 59 (1999), 113–31. (I should emphasize that I am using the term 'basic rule' here in a different sense from the sense that I gave the term in that work.)

14 This is, of course, only *one* way of distinguishing between what "may be a matter of luck" and what "lies within one's control". There are many other ways of understanding this distinction on which non-basic actions may sometimes be "within one's control", or one's performance of a basic action may sometimes be "a matter of luck".

rule, and an opportunity to follow that rule arises, then whether or not one follows that basic rule lies within one's direct control; it cannot be a matter of mere luck. So if rationality is a matter of following basic rules, this could explain why rationality lies within one's direct control, and cannot be a matter of mere luck.

According to my definition, a "basic rule" is a rule that one can follow directly, not by means of following any other rule. But what would it be to follow one rule by means of following another rule? For example, consider the rule: "Add salt when the water starts boiling". If one follows this rule, then one's adding the salt is explained by, or is a response to, the fact that the water is boiling. However, it may be that the process whereby one's action of adding the salt is explained by, or responds to, the fact that the water is boiling can itself be analysed, even at the folk-psychological level of explanation, into a series of *sub-processes*. For example, perhaps the proximate explanation of one's attempt to add the salt is not the *fact* that the water is boiling, but rather one's *belief* that the water is boiling. Similarly, perhaps the proximate explanation of one's belief that the water is boiling is not the *fact* that the water is boiling, but one's having an *experience* that represents the water as boiling. In forming this belief in response to having this experience, one is following a rule. Specifically, one is following a rule that permits one to form the belief that the water is boiling, when one has an experience that represents the water as boiling (and no special reason to distrust one's experience in the circumstances). Similarly, when one attempts to add the salt, in response to one's forming the belief that the water is boiling, one is following some other rule (or set of rules). In that case, the process of one's following the rule "Add salt when the water starts boiling" is constituted by a series of sub-processes, including (among other things) the processes of one's following those other rules. If it is only by means of such a series of sub-processes that one can follow the rule, then the rule "Add salt when the water starts boiling" is not a basic rule. One can follow this rule only by means of following other rules.

On the other hand, for example, consider the rule "Believe p, if one has an experience as of p's being the case, and no special reason to distrust one's experiences in the circumstances". If this is a "basic" rule, then one can follow this rule "directly". When one follows this rule directly, the process of one's following this rule cannot be analysed, at the folk-psychological level of explanation, into a series of sub-processes that include one's following any other rule. At this level of explanation, one's having an experience as of p's being the case is (at least part of) the *proximate explanation* of one's coming

to believe p.[15] There are no intervening steps that can be captured at this folk-psychological level of explanation.[16]

When a correct folk-psychological explanation of a belief revision includes all the intervening steps that can be captured at the folk-psychological level, let us call it a "fully-articulated" explanation. If one directly follows a basic rule, which permits one to revise one's beliefs in a certain way whenever one is in a certain condition, then a fully-articulated explanation of that belief revision will identify one's being in that condition as (at least part of) the proximate explanation of that belief revision.

It is important that this claim only concerns the *personal, folk-psychological* level of explanation. At a "subpersonal" level of explanation, it may well be that the process of one's directly following this rule can be analysed into numerous sub-processes, perhaps involving various subpersonal modules' computing various algorithms. But this is not the sort of explanation that we are concerned with here. We are concerned with explanations that have the following two features. First, these explanations are at the *personal, mental* level: what is explained is a mental fact about a person as a whole—such as the person's having or forming a certain mental state, like a belief or an intention, of the sort that are referred to in everyday folk-psychological discourse. Moreover, this fact is explained by reference to other states of the person as a whole; these explanations do not refer to states of subpersonal mechanisms or modules in the brain or anything of that sort. Second, these explanations make the person's having or forming that mental state *intuitively intelligible* or *unsurprising*. For example, the fact that John decided to go to the florist's shop this morning is made intuitively intelligible by the fact that he wanted to buy some flowers, and believed that the best way to do this was to go to the florist's shop this morning. On the other hand, John's going to the florist's shop this morning is *not* made intuitively intelligible by the fact that he wanted to see the new Steven Spielberg movie, and believed that the best way to do that is to go to the cinema in the

---

15 I am assuming here that if the process of one's revising one's beliefs through following a rule can be analysed (at the folk-psychological level of explanation) into a series of sub-processes, then at least one member of this series of sub-processes must be the process of one's following some other rule. This is why the process of one's following a rule "directly" cannot be analysed (at the folk-psychological level of explanation) into any sub-processes at all.

16 I should note that I am using the terms 'explain' and 'explanation' in a systematically ambiguous way. When I speak of giving an "explanation" of why internalism is true, I have in mind a *philosophical* explanation of a necessary truth. When I speak of a "psychological explanation" of a belief revision, I have in mind an *empirical* explanation of a contingent truth. Finally, I also use the term 'explanation' to refer both to the explanatory accounts that are given by theorists, and to a fact which those theorists could correctly cite as the *explanans* of whatever they are trying to explain. This ambiguity should cause no confusion in context.

evening.[17] I shall refer to explanations that have these two features as "folk-psychological explanations".

4 The Proximate Explanation of a Belief Revision

In this section, I shall argue that, whenever a thinker revises her beliefs through following a rule, a fully-articulated folk-psychological explanation of that belief revision will always identify the *proximate explanation* of that belief revision with an "internal fact" about the thinker's mental states. As I shall define it, the term 'internal fact' applies to any fact that supervenes purely on the thinker's "non-factive" mental states, and also to any fact about the explanatory relations in which such internal facts stand to each other. (The defining mark of a "factive" mental state, such as *knowing* or *seeing* that p is the case, is that it must consist in standing in some relation to a *true* proposition. If one knows or sees that p is the case, then p must actually be the case.)

According to this definition, the fact that the thinker is in a certain brain state does *not* count as an "internal fact" about the thinker's mental states. This fact does not supervene purely on the thinker's non-factive mental states (there are possible worlds w_1 and w_2 such that the thinker has exactly the same non-factive mental states in both w_1 and w_2, but in w_1 he is in the brain state in question while in w_2 he is not). This fact is also not a fact about the explanatory relations between "internal facts". As I shall put it, this fact is "external to the thinker's mind", or for short, an "external fact".

Some philosophers may object that it is surely an *empirical* question whether it is ever correct to explain a belief revision directly on the basis of an "external" fact of this sort. Certainly, it is an empirical question what facts can explain a belief revision. But there may still be certain philosophical limits on which correct explanations of a belief revision count as "fully articulated folk-psychological explanations" of the relevant sort.

Thus, I am not denying that the formation of a belief can *ever* be explained in terms of external facts. There may certainly be correct *non*-folk-psychological explanations that identify such an external fact as the proximate explanation of why one formed a belief. For example, scientists might discover that a certain brain state always causes the thinker to believe that he is about to die. But this would not be a folk-psychological explanation of the

---

[17] There are many theories about what it is to make someone's having or forming a certain mental state "intuitively intelligible" in this way. On some theories, it is a matter of explaining the mental state in accordance with a certain tacitly known folk-psychological *theory*. On other theories, it is a matter of *Verstehen*—that is, imaginative projection into, or simulation of, the person's point of view. I will remain neutral between these different theories here. I shall simply have to rely on the reader's having an intuitive sense of when explanations succeed in making someone's having or forming a certain mental state intuitively intelligible or unsurprising.

relevant sort. It would not make it intuitively intelligible or unsurprising that the thinker found this belief persuasive or compelling in the circumstances. From a folk-psychological perspective, the belief would still seem opaque and hard to understand.

I am also not denying that it could be the case, for example, that my friend Matthew's coming to believe that I once lived in Malaysia is explained by the external fact that I *told him* that I once lived in Malaysia. This explanation may be quite correct. It is just not a "fully-articulated" explanation. Intuitively, it seems, if this is a correct folk-psychological explanation, there must also be a more detailed correct folk-psychological explanation, in which the link between my telling Matthew that I once lived in Malaysia and his coming to believe that I once lived in Malaysia is mediated by intervening internal facts about his mental states. Perhaps, for example, in this more detailed explanation, Matthew's coming to believe that I once lived in Malaysia is directly explained by his having the *belief* that I told him that I once lived in Malaysia (along with the fact that he has no mental states that give him any reason to doubt that my assertion is true). This belief (that I told him that I once lived in Malaysia) is itself explained by his having an *experience* as of my telling him that I once lived in Malaysia, which is in turn explained by my actually telling him that I once lived in Malaysia.

Suppose that I claim that someone's forming a certain belief is explained by a certain external fact, in a context in which it is unclear how there could be any more detailed correct explanation in which the link between that external fact and the formation of the belief is mediated by any intervening internal facts about the thinker's mental states. For example, suppose that I say, "I once lived in Malaysia, so Vladimir Putin believes that I once lived in Malaysia". It would be natural for you to reply, "But how does Putin know anything about you at all? Did you meet him and talk about your childhood? Did he investigate you while he worked for the KGB? Or what?" In asking these questions, you seem to reveal that you would not accept this explanation unless it is plausible to you that this link, between the fact that I once lived in Malaysia and Putin's believing that I once lived in Malaysia, is mediated by intervening internal facts about Putin's mental states.

This point applies even to perceptual beliefs. Suppose that I claim "Sarah believes that the flowers in front of her are pink because the flowers *are* pink", while simultaneously claiming that this link, between the flowers' being pink and Sarah's believing that the flowers are pink, is not mediated by any intervening internal facts about Sarah's mental states. If these claims are correct, then either Sarah has no experience that represents the flowers in any way, or else, if she has such an experience, it makes absolutely no difference

to whether or not she forms this belief.[18] But then how can this explanation make this belief intuitively intelligible or unsurprising? How exactly does the mere fact that the flowers are pink make it persuasive or compelling for Sarah to form precisely this belief, rather than some other belief, or indeed any belief at all? If we want to make it intuitively intelligible why Sarah found it compelling to form precisely this belief, the most plausible answer is surely to say something like: "Sarah found it compelling to form this belief because it *looked to her* as though those flowers were pink".

Some philosophers might concede that it would be strange or surprising if Sarah formed the belief that the flowers are pink, because of the external fact that the flowers really are pink, without having any experience that represented the flowers in any way. But they still might deny that the experience is an "intervening mental state" mediating between that external fact and that belief. For example, these philosophers might suggest that the experience and the belief are independent effects of an external common cause. But if that suggestion were correct, then we could give a correct explanation of the belief purely by appealing to this external common cause, even if (for some unusual reason) no experience of the relevant kind ever occurred. As I have just argued, however, this explanation could not be a correct *folk-psychological* explanation of the belief. It would fail to make the belief intuitively intelligible, as something that it was intuitively unsurprising that the believer found persuasive or compelling in the circumstances.

Alternatively, some philosophers might suggest that the experience just *is* the perceptual belief, so that the belief cannot be *explained by* the occurrence of the experience. But this suggestion can also be ruled out: I am understanding the explanandum here as consisting in the person's having or forming a mental state of a certain *type*, not in a particular mental state *token*. The mental state type *believing that the flowers are pink* is undoubtedly distinct from the type *having an experience that represents the flowers as pink*. One might believe that the flowers are pink without having any experience that represents the flowers as pink; and vice versa. (For example, a blindfolded person might feel some flowers, and irrationally believe for no reason that they are pink. Or someone might have an experience that represents the flowers as pink, but distrust her own experience for some reason and so refuse to believe that the flowers are pink.)

In general, then, it seems that an explanation of a belief revision that appeals to an external fact can be a correct folk-psychological explanation only if there is also a "fully-articulated" explanation in which the link

[18] It is hard to imagine what this belief could be like, especially if we suppose that Sarah does not have any experience that represents the flowers in any way. But perhaps "blindsight" cases would be an example. It is certainly not at all clear that it is rational to form beliefs on the basis of "blind-sight" in the same way as it is to form perceptual beliefs on the basis of ordinary experience.

between that external fact and that belief revision is mediated by intervening internal facts about the believer's mental states. So, in any "fully-articulated" explanation, the proximate explanation of the belief revision is not an external fact, but some internal fact about the believer's mental states. Typically, this proximate explanation involves the experiences, apparent memories, intuitions or beliefs that are the *reasons* for which the believer revised his beliefs in that way, along with the absence from his set of mental states of certain sorts of defeating or countervailing reasons. It is striking how sharply beliefs contrast with experiences on this point. The proximate folk-psychological explanation of an experience typically *does* involve an external fact: "The body was lying there right in front of her, in broad daylight, and her eyes were wide open, so of course she saw it". These explanations never appeal to any "reason" for which one has that experience.

One might object that the arguments given so far only show that the proximate explanation of a belief revision must involve some fact about the believer's *mental states*. It does not show that it must be an *internal* fact about the believer's mental states. As I defined the term above, an "internal fact" is either a fact that supervenes purely on the thinker's "non-factive" mental states, or else a fact about the explanatory relations between such internal facts. But why cannot the proximate explanation of a belief revision sometimes involve "factive mental states", such as the state of *knowing that p is the case*, or *seeing that p is the case*?[19]

In fact, it seems that such "factive states" cannot figure in the proximate explanations of belief revisions, in any correct fully-articulated folk-psychological explanations. Suppose that we want to explain why a thinker comes to believe p. One candidate explanation identifies the proximate explanation of the thinker's coming to believe p as the fact that the thinker *knows* q; another candidate explanation identifies this proximate explanation as the fact that she *believes* q. In this case, if either explanation is correct, it is the second explanation, not the first. It is highly plausible that the thinker's knowing q is *partially constituted* by the thinker's believing q; and if the thinker had merely believed q, and not known q, she would still have come to believe p, in exactly the same way that we are trying to explain. So, the thinker's knowing q has the effect of producing the belief in p only because her knowing q is partially constituted by her believing q. This seems to show that it is the thinker's believing q, not her knowing q, that really explains her having the belief in p that we are trying to explain.[20]

---

[19] One externalist who insists that these "factive mental states" are among the mental states that determine whether or not a belief is rational or justified is Williamson; see especially *Knowledge and its Limits*, chapter 9. Compare also John McDowell, "Knowledge and the Internal", *Philosophy and Phenomenological Research* 55 (1995), 877–93.

[20] Objection: What if there is a *time lag* between the thinker's being in the relevant condition (say, believing q and considering the question of whether p is the case) and her

This argument is an application of a plausible general principle about explanation. If one fact is partially constituted by a second,[21] and a certain effect would still have been produced even if the second fact had obtained while the first fact had not, then if either fact explains that effect, it is the second fact rather than the first. The first fact contains elements that are irrelevant to explaining the effect: it is the second fact that really does the work in explaining that effect.

I am not denying that knowledge *ever* plays a role in folk-psychological explanations. As Williamson has recently argued, knowledge does seem to play such a role in the explanation of certain *actions*.[22] For example, perhaps I keep on digging because I *know* that this mine contains gold. Believing, even truly believing, that it contains gold would not have been enough; for then I might have inferred this belief from a lemma whose falsity I might easily have discovered while digging, in which case I would have abandoned the belief and stopped digging. Here, however, the explanandum—my keeping on digging—consists in an agent's interacting with his environment in a certain way. It is only to be expected that the explanans—my knowing that the mine contains gold—will also consist in the agent's standing in a certain relation to his environment. This does not show that knowledge will figure in the explanation of an "internal" fact, such as the fact that a thinker *comes to believe p at time t*. An internal fact of this sort is surely more likely to have a correspondingly internal explanation.

---

responding to that condition by coming to believe p? Then it might not be true that if she had merely believed q and not known q, she would still have come to believe p. If her belief in q did not amount to knowledge, then she could easily have encountered evidence during that time lag that would have led her to abandon her belief in q, in which case she would not have come to believe p. Reply: The *proximate* explanation of the thinker's coming to believe p at a certain time t must surely involve the fact that the thinker believed q during a period of time leading up to t. This explanation cannot leave it open whether or not she still believes q at t. So, even if there is a time lag between her considering the question of whether p is the case and her coming to believe p, there is no such time lag between her believing q and coming to believe p.

21 This clause is important, to get round the objection that this counterfactual test will always lead one to prefer the most *disjunctive* explanations possible. The truth of a proposition is *not* "partially constituted" by the truth of a disjunction of which that proposition is a disjunct—whereas knowing p is partially constituted by believing p. This "plausible general principle about explanation" is analogous to a principle about causation that is defended by Stephen Yablo, "Cause and Essence", *Synthese* 93 (1992), 403–49, especially pp. 413–23, and "Wide Causation", *Philosophical Perspectives* 11 (1997), 251–81. Some closely related ideas about explanation are defended in Williamson, *Knowledge and its Limits*, pp. 80–88.

22 *Knowledge and its Limits*, pp. 60–64, 75–88. Of course, if we can give a non-circular definition of knowledge in terms of other folk-psychological notions—for example, if knowledge can be defined as a rational belief that is in a certain sense "reliable", as I believe—then knowledge would not play an *indispensable* role in any of these explanations. But I cannot go into this question here.

In general, the overall effect of the principle about explanation that I am appealing to here is that in any correct explanation there must be a certain sort of *proportionality* between the explanandum and the explanans. The explanans must be sufficient in the circumstances to produce the explanandum; but it also must not contain any irrelevant elements that could be stripped away without making it any less sufficient to produce the explanandum. For this reason, we need not worry that this principle will lead to the conclusion that the *content* of one's mental states cannot be both explanatorily relevant and determined, in part, by one's relations to one's external environment. The explanandum, in all the cases that we are concerned with, is the formation of a belief with a certain content. If the content of the belief that figures in the explanandum is itself determined by the thinker's relations to her environment, it is only to be expected that the explanation of this belief will involve mental states whose content is also determined by the thinker's relations to her environment. The trouble with the idea that one's forming a certain belief may be explained by what one knows is not that what one knows depends on the environment at all. The trouble is that what one knows is *too* dependent on the environment to give a suitably proportional explanation of one's forming this belief. This is shown by the fact that one's knowing q is partially constituted by one's believing q, and yet if one had merely believed q and not known q, one would still have formed the belief in p.

A parallel argument can be given for other factive states as well. Suppose that we want to explain why a thinker comes to believe p. According to the first candidate explanation, the thinker's coming to believe p is explained by the fact that she *sees* that p is the case (and has no special reason to distrust her senses in the circumstances). According to the second candidate explanation, the thinker's coming to believe p is explained by the fact that she has an *experience* as of p's being the case (and has no special reason to distrust her experience in the circumstances). Assuming that the thinker's seeing that p is the case is partially constituted by her having an experience as of p's being the case,[23] then if either explanation is correct, it will be the second explanation, not the first. If the thinker had merely had an experience as of

[23] This assumption is denied by those who hold a "disjunctive" view of experience. For a classic statement of the disjunctive view, see John McDowell, "Criteria, Defeasibility and Knowledge", *Proceedings of the British Academy* 68 (1982), 455–79. For criticism of some of the arguments that are used to support this disjunctive view, see Alan Millar, "The Idea of Experience", *Proceedings of the Aristotelian Society* 96 (1996), 75–90. The main argument against the disjunctive view is the "Argument from Hallucination"—that is, the argument that it is only if there is a common factor in veridical perception and hallucination that we can explain certain subjectively seamless transitions between perception and hallucination, as well as the fact that the perception and the hallucination both incline one to form such strikingly similar beliefs. For a powerful restatement of the Argument from Hallucination, see Mark Johnston, "The Obscure Object of Hallucination", forthcoming in *Philosophical Studies* (2002).

p's being the case, and never actually *seen* that p was the case, then (so long as the thinker still had no special reason to distrust her experiences in the circumstances) the thinker would still have come to believe p. So it is the fact that the thinker has an experience as of p's being the case, not the fact that she sees that p is the case, that really does the work in explaining why the thinker comes to believe p. It seems then that correct fully-articulated folk-psychological explanations will always identify the proximate explanation of a belief revision with an "internal fact" about the thinker's "non-factive" mental states.

Suppose that one directly follows a basic rule that permits one to revise one's beliefs in a certain way, whenever one is in a certain condition. Then, as I explained in §3, a correct fully-articulated explanation will identify the fact that one is in that condition as at least part of the *proximate explanation* of that belief revision. So, the fact that one is in this condition must itself be an "internal fact" about one's mental states. In general, following such basic rules always involves revising one's beliefs in response to such internal facts. This is not to say that it is impossible to follow a rule that permits one to revise one's beliefs in a certain way whenever a certain *external* fact obtains. It may be quite possible, for example, to follow the rule that permits one to believe p whenever one can *see* that p is the case. But this rule cannot be a "basic rule". If one follows this rule, one does so *by means of* following some basic rule, such as the rule that permits one to believe p whenever one has a visual experience as of p's being the case (and no reason to distrust one's experience in the circumstances).

Suppose that one revises one's beliefs by directly following this basic rule: one comes to believe p in response to the internal fact that one has an experience as of p's being the case, and no reason to distrust one's experience in the circumstances. Then the fact that one is directly following this rule is itself a fact about a certain explanatory relation that holds between this internal fact and one's coming to believe p. That is, the fact that one is directly following this rule is itself a fact about the explanatory relations in which one internal fact stands to another. So, given my definition of the term, the fact that one is directly following the rule is itself an "internal fact". In general, for any set of basic rules, the fact that one is directly following those rules will always count as an "internal fact" of this sort.

5 Belief Internalism and Rule Internalism

I have not yet explained why internalism is true. Even if the fact that one is directly following certain basic rules is always an internal fact of this sort, it could still be that the fact that it rationally "makes sense" for one to conform to these rules is *not* an internal fact of this sort. For example, it could be that what makes it the case that it "makes sense" for one to conform to these rules

is the fact that these rules are highly reliable at yielding true beliefs and avoiding false beliefs. As I shall put it, this position would combine "belief internalism" with "rule externalism".[24]

The closest parallel in the literature to my argument for belief internalism is the "refutation of belief externalism" that is given by John Pollock and Joseph Cruz.[25] After giving their refutation of belief externalism, Pollock and Cruz turn to "rule externalism". Specifically, they consider the claim that the basic rules that it rationally makes sense for us to conform to are those rules that are most reliable at reaching the truth. They understand this as the quite general claim that the basic rules that it rationally makes sense for us to conform to are *all* and *only* those basic rules that are reliable in this way—including both rules that we know to be reliable, and rules that we do not know to be reliable in this way.

Pollock and Cruz first point out that if this claim is to address the epistemological issues that concern us, this claim must be a *recommendation* about which rules to conform to. Specifically, it must be the general recommendation that we should conform to all and only reliable basic rules—in effect, the recommendation to reason in the most reliable way. But they object that this "is not a recommendation anyone could follow". Their reason is that "we can only alter our reasoning in response to facts about reliability if we are apprised of those facts" (*op. cit.*, p. 140).

Here, Pollock and Cruz seem to infer from the premiss 'We can only alter our reasoning in response to facts about reliability if we are apprised of those facts' to the conclusion 'No one can follow the recommendation to reason in the most reliable way'. If this inference were valid, then we could also infer from the premiss 'We can only add salt to the water in response to the fact that the water has started boiling if we are apprised of the fact that the water has started boiling' to the conclusion 'No one can follow the recommendation

---

[24] One quick way to dismiss rule externalism would be to claim that the fact that it makes sense for one to conform to a certain rule is always a necessary, non-contingent fact; then this fact would trivially supervene on internal facts. But this claim is implausible. The fact that it makes sense for *me* to conform to a rule must surely depend on contingent facts about me, such as what capacities I have or what evidence I have encountered.

[25] See Cruz and Pollock, *Contemporary Theories of Knowledge*, 2nd edition (Lanham, Maryland: Rowman & Littlefield, 1999), pp. 130–37. Their argument goes roughly as follows. Epistemic norms are "procedural norms" that can be "internalized". When norms are internalized, this enables "our cognitive system to follow them in an automatic way without our having to think about them". So, the circumstance-types in response to which these norms tell us to do something must be "directly accessible to our system of cognitive processing"; that is, "our cognitive system must be able to access them without our first having to make a *judgment* about whether we are in circumstances of that type". States that are in this way "directly accessible to our cognitive system" are what Cruz and Pollock call "internal states". The problem with this argument, as I see it, is that not enough is said about what it means for "our cognitive system" to "follow a norm in an automatic way", or to "access" a circumstance-type. This makes their argument somewhat hard to evaluate.

to add salt when the water starts boiling'. But that inference *cannot* be valid. Even if the premiss is true, it is obviously possible to follow the recommendation to add salt when the water starts boiling.

Pollock and Cruz seem to be assuming that a "recommendation that someone could follow" must be a recommendation that we can *always* follow *whenever* it applies to us. But how many recommendations are there of which that is true? Take the simplest of logical precepts: "From '*p* & *q*' infer *p*". We are not *always* able to follow this recommendation: some conjunctions are too complex to be recognized as such; or, more simply, we might suddenly die, or go insane, or fall asleep, before we have completed the inference; and so on. The only general recommendations that we can always follow, whenever they apply to us, are recommendations that are specifically restricted to cases in which we are able to follow them—for example, "From '*p* & *q*' infer *p*, whenever you are able to follow this precept". But recommendations of this sort might be externalist recommendations, such as "Form your beliefs by reliable methods, whenever you are able to follow this recommendation". So the idea of a recommendation that we are *always* able to follow does not support internalism. To support internalism, we need a different idea—roughly, the idea of a recommendation that we can follow *directly*, not by means of following any other recommendation.

According to my proposal, a belief revision is rational just in case it results from one's directly following rules that it "makes sense" for one to conform to, in order to pursue the relevant aim (say, the aim of believing the proposition in question if and only if it is true). To say that it "makes sense" to conform to a certain rule in order to pursue a certain aim is to *recommend* that rule as a means to that aim. As I remarked earlier, it is not immediately clear what sort of recommendation this is. I shall now propose that it must be understood as a recommendation that can be followed "directly" in this sense.

Consider a rule that it rationally makes sense for one to conform to, which permits one to revise one's beliefs in a certain way whenever one is in a certain related condition. As I argued above, one would not count as "following" this rule if it were simply a fluke that one revises one's beliefs in that way at the same time as being in the relevant condition. It must also be the case that one revises one's beliefs in the relevant way *in response to* one's being in the relevant condition. That is, one must revise one's beliefs in that way precisely *because* one is in the relevant condition

In order to count as "following" this rule, then, it is necessary that one's belief revision should be a "response" to one's being in the relevant condition—necessary, but, it seems, not sufficient. Suppose that you do revise your beliefs in this way, in response to being in the relevant condition. It could still be the case that your responding to this condition by revising your

beliefs in this way is just an uncontrollable compulsion, implanted into your brain by a manipulative neuroscientist, unconnected with any more general ability to revise your beliefs in rational ways. In that case, it might be a pure fluke that you are conforming to a rule that it rationally "makes sense" for you to conform to. But then you would not be genuinely *following* the rule; and the belief revision that you make in conforming to the rule in this way would not count as rational.

For this reason, I propose a further necessary condition for following the rule. To count as following the rule, one must revise one's beliefs in this way on this occasion, not only because one is in the relevant condition, but *also* because this is a condition in which revising one's beliefs in this way is something that it rationally makes sense for one to do. That is, one must revise one's beliefs in this way, not only because one is in the relevant condition, but also because it makes sense for one to conform to the relevant rule.

I am not proposing here that whenever one follows this rule, one must actually *believe* that it "makes sense" for one to conform to the rule. Such a higher-order belief is neither necessary nor sufficient for following the rule. It is not necessary because unsophisticated thinkers (such as young children) may revise their beliefs through following rules, even though they do not even have the concept of a rule that it "makes sense" for them to conform to. It is not sufficient because even if one holds such a higher-order belief, one's disposition to conform to the rule could be wholly independent of this belief, and so it could still be an uncontrollable compulsion, unconnected with any more general ability to reason in rational ways. Moreover, imposing such a higher-order belief requirement would also generate a *regress* if this higher-order belief must itself result from the thinker's following a rule.

What I am proposing, instead, is that whenever one follows the rule, the *fact* that it rationally makes sense for one to conform to the rule must itself be part of the explanation of the belief revision in question. One conforms to the rule on this occasion—that is, one revises one's beliefs in this way on this occasion—at least in part *because* this really is a rule that it makes sense for one to conform to.

Even with this further necessary condition, however, we still do not have a sufficient condition for following a rule. After all, the manipulative neuroscientist might be motivated by benevolence. He might have implanted this uncontrollable compulsion into your brain precisely *because* this compulsion will lead you to conform to a rule that it makes sense for you to conform to. But this would still not make it the case that you are genuinely following the rule. Clearly, you are not following this rule "indirectly", because your conforming to this rule on this occasion is not something that you do by means of following any *other* rules. But why isn't this a case of your following the rule "directly" (that is, not by means of following other rules)?

The proposals that I have made above suggest that the answer is this. In this case, even though your conforming to this rule on this occasion is explained, at least in part, by the fact that this is a rule that it rationally makes sense for you to conform to, this fact does not *directly* explain your conforming to the rule. The link between this fact and your conforming to the rule on this occasion is mediated by the intervention of the neuroscientist. That is, this fact is not part of the *proximate* explanation of the belief revision in question.[26]

If this proposal is correct, then we can draw the following general conclusion. Whenever one revises one's beliefs by "directly" following a rule that it makes sense for one to conform to, the fact that it makes sense for one to conform to the rule must be part of the proximate explanation of the belief revision in question. As I argued in §4, however, the only facts that can form part of the proximate explanation of a belief revision, in a fully-articulated folk-psychological explanation, are *internal facts* about one's (non-factive) mental states. Since the fact that it makes sense for one to conform to this rule is part of the proximate explanation of this belief revision, this fact must also be such an internal fact about one's mental states.

Thus, for example, the fact that conforming to this rule is a reliable way of getting to the truth is not an internal fact of this sort. This fact does not supervene on one's non-factive mental states: there are possible worlds w_1 and w_2, such that one has exactly the same non-factive mental states in both worlds, but in w_1 the rule is reliable while in w_2 it is not. Moreover, the fact that the rule is reliable is also not a fact about the explanatory relations in which such internal facts stand to each other. So the fact that the rule is reliable cannot form part of the *proximate* explanation of a belief revision. Thus, the fact that it makes sense for one to conform to this rule cannot be identified with the fact that the rule is reliable in this way.

Clearly, it is a highly challenging question exactly what constitutes the fact that it rationally makes sense for one to conform to a certain basic rule. Fortunately, we do not need to answer this question here. A basic rule is a rule that one can follow directly. So, according to my proposal, the fact that it makes sense for one to follow such a basic rule must be capable of being part of the proximate explanation of a belief revision. Hence, it must be an internal fact about the thinker's non-factive mental states.

---

[26] How is it possible for the fact that it rationally makes sense for one to conform to a certain rule to be part of the proximate explanation of a belief revision? The answer might be roughly as follows. In any such case, the belief revision results from the activation, not only of one's disposition to conform to this rule, but also of one's more general disposition to conform to rules that it rationally makes sense for one to conform to—where the activation of this general disposition on this occasion cannot be analysed into any series of subprocesses at the folk-psychological level of explanation. But I cannot further investigate this question here.

At the end of §4, I argued that for any set of basic rules, the fact that one is directly following these rules is always an internal fact of this sort. I have now argued that it is also an internal fact of this sort whether or not it rationally makes sense for one to conform to these rules. Thus, we have now have an explanation, not just of why "belief internalism" is true, but also of why "rule internalism" is true.

In this way, then, internalism can be explained. The "internal facts" on which the rationality of a belief revision supervenes are either facts that supervene on one's non-factive mental states, or facts about the explanatory relations in which such internal facts stand to each other. The rationality of belief revisions supervenes on such internal facts for the following reason. The rationality of a belief revision depends on the basic rules that one was directly following in making that belief revision, and so on the proximate explanations of the belief revisions that one made in following those rules; and as I have argued, the proximate explanation of a belief revision—the fact to which that belief revision most directly and immediately responds—is always an internal fact of this sort.[27]

[27] Earlier versions of this paper were presented to the Philosophy Department at Stanford University and to the 2001 Rutgers Epistemology Conference. I am most grateful to the members of both audiences, and also to Stephen Yablo, Timothy Williamson, David Velleman, James Pryor, John Gibbons, and Alexander Bird, for many helpful comments.