Self-knowledge and the Paradox of Belief Revision

Giovanni Merlo¹

Accepted: 22 March 2021 / Published online: 6 April 2021
© The Author(s) 2021

Abstract

To qualify as a fully rational agent, one must be able rationally to revise one's beliefs in the light of new evidence. This requires, not only that one revise one's beliefs in the right way, but also that one do so as a result of appreciating the evidence on the basis of which one is changing one's mind. However, the very nature of belief seems to pose an obstacle to the possibility of satisfying this requirement – for, insofar as one believes that p , any evidence that not- p will strike one as misleading and, on the face of it, believing that a certain piece of evidence is misleading is incompatible with appreciating the fact that such evidence should bear on the question at hand. Call this the 'Paradox of Belief Revision'. This paper introduces the Paradox of Belief Revision, compares it with Kripke's Dogmatism Paradox, and suggests that we may be able to see a way out of the former if we assume that rational agents are systematically aware of their own beliefs as beliefs they have.

Keywords Self-knowledge · Rationality · Belief revision · Dogmatism paradox

Self-knowledge – the distinctively 'first-personal' knowledge that each of us has of his or her own mental states – seems to differ from other kinds of knowledge in several philosophically interesting ways. One of these concerns the apparent connection between a subject's possessing self-knowledge (or, at least, certain selected pieces of self-knowledge) and that subject's qualifying as rational. On the face of it, much knowledge is inessential to rationality. Being rational requires such things as forming one's beliefs in a rationally appropriate way and avoiding certain rationally inappropriate combinations of beliefs – it does not require having specific beliefs, let alone *knowledge*, about this or that subject matter. But *self*-knowledge would seem to provide an exception to this general rule. Arguably, a subject who knew nothing at all about his or her own beliefs, except what could be known equally well by others on the basis of 'third-person' evidence, would be rationally – not just epistemically – deficient. Conversely, it seems to be a mark of rational

✉ Giovanni Merlo
giovanni.merlo@unige.ch

¹ Philosophy Department, University of Geneva, Rue de Candolle 2, Room 401, 1205 Genève, Switzerland

belief possession that, whenever one rationally believes something, one has (or is in a position to have) some distinctively ‘first-personal’ knowledge of the fact that one believes it. These points, though by no means uncontroversial, are central to so-called ‘constitutivist’ accounts of self-knowledge and, to the extent that they apply to self-knowledge of one’s own beliefs, it is plausible to think that they should carry over to self-knowledge of other mental attitudes, including, for instance, desires and intentions.¹

Sticking to the case of belief, one way to bring out the apparent connection between self-knowledge and rationality is to try to construct a case in favour of the following *Reflective Rationality Thesis*:

Reflective Rationality Thesis: Rational belief revision requires the ability to form true second-order beliefs about one’s own beliefs.

This thesis affirms that any subject who can rationally revise his or her own beliefs must have the ability to form true second-order beliefs about them – or *self-reflection*, as I am going to refer to this ability hereafter. Now, having self-reflection is a precondition for possessing self-knowledge of one’s own beliefs. And rationally revising one’s own beliefs is something that any agent should be capable of doing, if he or she is to qualify as rational. Thus, if the Reflective Rationality Thesis holds true, a certain indirect connection holds between rationality and self-knowledge of one’s own beliefs. A direct connection could be established, on the same basis, by showing that the kind of true second-order beliefs required by rational belief revision must qualify as knowledge and be distinctively first-personal – or, alternatively, that the general ability to form true second-order beliefs about one’s own beliefs is somehow derivative on a specific ability to acquire self-knowledge.

Some may be tempted to take the truth of the Reflective Rationality Thesis for granted. Rationally revising one’s beliefs is a matter of adjusting one’s beliefs to one’s evidence. And it may be thought that, in order to adjust one’s beliefs to one’s evidence, one must have knowledge of both – in the same way that, in order to adjust one’s expenses to one’s income, one must know both how much one spends and how much one earns. But, of course, the analogy is open to all sorts of criticism. Since we patently do *not* stand in the same relationship to our beliefs as we do to our money, there is no obvious reason to think that a necessary condition for managing the latter should be a necessary condition for ‘managing’ the former. Unsurprisingly, then, defenders of the Reflective Rationality Thesis have been under pressure to provide independent arguments in favour of their view. Shoemaker (1988), Burge (1996, 99–100) and Gallois (1996, 75–80) can all be seen as making different contributions in this direction.

This paper engages the above dialectic in a somewhat indirect fashion. My goal is to present a certain paradox attending the idea of rational belief revision and to suggest that we may be able to see a way out of the paradox if we assume that rational agents are systematically aware of their own beliefs as beliefs they have. This suggestion, if correct, would provide grist to the mill of the Reflective Rationality Thesis, but only in a roundabout and limited way. For one thing, I will not attempt to argue that we can *only* solve the *Paradox of Belief Revision* (as I am going to call it) by assuming that rational agents are self-reflective. It is possible that there are other ways of solving the

¹ See, among others, Wright (1987), Shoemaker (1988), Burge (1996), Bilgrami (1998), Moran (2001). For criticism, see Reed (2010).

paradox, and that some of these do not involve any commitment to the truth of the Reflective Rationality Thesis. Moreover, what I will be able to offer is nothing more than the blueprint of a solution to the paradox. My account of how self-reflection contributes to make rational belief revision possible has some significant loose ends, which I will point out in due course.

The paper is divided into four sections. In § 1 and 2, I will introduce the Paradox of Belief Revision and distinguish two possible approaches to tackling it. In § 3, I will compare the Paradox of Belief Revision to Kripke's Dogmatism Paradox. The comparison will allow us to identify which of the two approaches distinguished in § 2 should be pursued if we are to solve the two paradoxes in a satisfactory way. Finally, in § 4, I will explain how the assumption that rational agents are systematically aware of their own beliefs as beliefs they have might help us implement the needed approach with respect to the Paradox of Belief Revision.

1 The Paradox of Belief Revision

The experience of changing one's mind is one of the most common and familiar. We change our mind all the time, for all sorts of reason. I leave my place with the intention of going to my office. I run into a friend, and decide to go to a café, instead. I believe that Hemingway is a good writer. I read one of his books, and decide that he is not so great, after all. I know that the Thirty Years' War ended in 1648. Time passes, I forget the exact date and form the wrong belief that it ended in 1649.

Among the changes of mind that do not simply 'happen' to us (as in the third example) but of which we are, or take ourselves to be, an active source (as in the first two examples), it is natural to think that an important distinction holds between the rational and the irrational ones. Take any case in which we actively change (or 'revise') our own beliefs. We would like to think that only some such cases are cases of *rational belief revision*, properly so termed. When I go from believing that $2 + 2 = 5$ to believing that $2 + 2 = 4$ via learning arithmetic, I rationally revise my belief that $2 + 2 = 5$. By contrast, when I go from believing that $2 + 2 = 5$ to believing that $2 + 2 = 4$ on the basis of random guessing, I revise my belief that $2 + 2 = 5$, but not in a rational way.²

Rational belief revision – or, at any rate, *evidence-based* rational belief revision, the type of rational belief revision on which I am going to focus from now on – seems to require at least two things. First, it seems to require that, upon acquiring some new evidence, one revise one's beliefs in a way that is, broadly speaking, *adequate* to the evidence one has acquired. Suppose I have strong evidence that Mr. Jones is the murderer and no evidence that he is not. If I go from believing that Mr. Jones is the murderer to believing that he is innocent based solely on the testimony of a shady friend of his whom I do not find generally trustworthy, I am *not* revising my beliefs in a rationally adequate way, because the change in my beliefs is not 'proportional' to the new evidence I have acquired. Call this the *Adequacy Condition*. Second, it seems that

² I shall reserve the term 'belief revision' for cases where the relevant change consists in the alteration or elimination of a pre-existing belief. The topic of belief acquisition falls outside the scope of my discussion. This restriction, while perhaps non-standard in philosophical discussions of belief revision, has some intuitive basis in our ordinary talk of 'changing one's mind', which presupposes a preformed opinion about the topic in question.

rational belief revision should result from some kind of *appreciation* of the fact that, given the evidence, things must be a certain way rather than another. The change of mind should not simply be proportional to the new evidence, it should also be ‘guided’ or ‘underpinned’ by the subject’s grasp of the significance of that evidence to the question at hand – for example, the significance of this or that testimony to the question whether Mr. Jones is the murderer. What is involved in this additional act of ‘appreciation’ is, of course, a difficult question, to which I shall come back in the next section. For the moment, let me refer to the requirement that rational belief revision should result from this yet-to-be-clarified appreciation of the new evidence as the *Appreciation Condition*.

I do not have any conclusive argument that rational belief revision should satisfy both the Adequacy and the Appreciation Condition. However, reflection on the parallel case of rational belief possession provides at least initial reason to think that this is so. Rationally possessing the belief that Mr. Jones is the murderer requires, not only having some evidence that Mr. Jones is the murderer, but also *basing* one’s belief that Mr. Jones is the murderer on that evidence – or, in the terminology of ‘reasons’, possessing the belief *for* the relevant reasons.³ It is unclear why an analogous requirement should not hold in the case of rational belief revision. A change of mind that adequately reflected the new evidence but did not result *from* that new evidence in the right way would not seem to be rational. And resulting from the new evidence ‘in the right way’ would seem to mean, among other things, resulting from a (yet-to-be-clarified) appreciation, by the subject, of that evidence.⁴

Now, the paradox I want to present makes trouble for the idea that we can rationally revise our own beliefs precisely by questioning the possibility of satisfying the Appreciation Condition. In a nutshell, the paradox could be put as follows: to the extent that one believes that p , one will take it for a fact that p , which means that one will regard evidence that not- p as misleading. But if one regards any evidence that not- p as misleading, one will not be able to appreciate the significance of any such evidence to the question whether p . Consequently, believing that p makes it impossible for the person who has the belief to revise it in a rational way.

To arrive at a more careful formulation of the paradox, it is useful to make some stipulations. First, let us restrict our attention to rationally held beliefs (the specification will often be left implicit from now on). Second, let us say that a certain piece of evidence E constitutes (or is) ‘misleading evidence that p ’ if, and only if, (i) E supports the hypothesis that p and (ii) it is false that p . Finally, let us consider a rational subject S who possesses the concept of misleading evidence and is aware that (for any p) ‘ p ’ and ‘This piece of evidence supports the hypothesis that not- p ’ entail ‘This piece of evidence

³ This corresponds to the classical distinction, first introduced by Firth (1978), between ‘propositional’ and ‘doxastic’ warrant (or justification). Cf. also Davidson’s (1963) closely related point that an agent who does something and has a good reason for doing it may nevertheless fail to do it *for* that reason. Another relevant parallel is with the ‘Taking Condition’ on inference, according to which “inferring necessarily involves the thinker *taking* his premises to support his conclusion and drawing his conclusion *because* of that fact” (Boghossian 2014, 5).

⁴ I expect that critics of the Taking Condition (see footnote 3 above) will take issue with this suggestion (cf. Wright’s (2014, 34) complaint that doing something for a reason does not require having awareness of those reasons as reasons). The question is too complex to take up here. At any rate, later in this section I will argue that, even if we deny that the Appreciation Condition is a necessary requirement on rational belief revision, the problem posed by the Paradox of Belief Revision still arises, only in a different guise.

is misleading'. Here is how a proponent of the Paradox of Belief Revision may argue that, whenever S believes something, she will not be able rationally to revise her belief:

- (1) If S believes that p , then (insofar as S considers the matter carefully enough) S believes, of any piece of evidence E that S takes to be evidence that not- p , that E is misleading evidence;
- (2) If S believes, of any piece of evidence E that she takes to be evidence that not- p , that E is misleading evidence, S cannot satisfy the Appreciation Condition (with respect to any such piece of evidence);
- (3) Hence, if S believe that p (and considers the matter carefully enough) S cannot rationally revise her belief that p .

I take the conclusion of this argument to be paradoxical, for at least two reasons. First, one would have thought it possible for a subject like S to revise her beliefs in a rational way. But if the above argument goes through, this is not so. Second, one would have thought it possible for a subject like S to have rational beliefs in the first place. But one may plausibly argue that rational belief possession presupposes the possibility of rational belief revision. One does not believe that p in a rational way unless one is suitably responsive to considerations for and against its being the case that p , and the relevant kind of responsiveness is one that should be capable of manifesting itself in a rational change of mind. Therefore, the above argument risks to be a *reductio* of the very possibility that S has any rational beliefs. Since all we assumed about S was that S is rational, possesses the notion of misleading evidence and is aware of a certain pattern of entailment involving that notion, the paradox (if it is one) generalizes to all subjects sharing with S these basic features.⁵

Of course, (3) follows from (1) and (2) only on the assumption that rational belief revision is subject to the Appreciation Condition. Perhaps, at this point, some will urge that this condition embodies an overly demanding (or, at any rate, misguided) conception of rationality. On an alternative and less demanding conception, not all cases of rational belief revision require that the subject 'appreciate' the significance of the new evidence to the question at hand. Whether this response can be plausibly sustained depends on how one understands the yet-to-be-clared notion of appreciation – a topic to which I shall return shortly. But note that, even if we adopt the less demanding conception, we are not necessarily off the hook. For it may turn out that the less demanding conception is somehow dependent on the more demanding one. For example, it may be suggested that, when we think of this or that case of belief revision as rational, it is always because we assume that the subject (or, perhaps, an idealized counterpart of the subject) could in principle satisfy the Appreciation Condition, whether or not she in fact does so. To the

⁵ Since the paradox targets the notion of *evidence-based* rational belief revision, those who equate believing that p with having credence 1 that p (see, for instance, Clarke 2013) may be unimpressed by it. Given a standard approach to conditional credence, if belief is credence 1, rational belief revision cannot result from standard Bayesian conditionalization on the new evidence. So, in a sense, the view that belief is credence 1 already entails that evidence-based rational belief revision is impossible. I regard this as an undesirable consequence of the view that belief is credence 1. Indeed, even if I want to remain neutral on the relationship between belief and credences, I think one might see the Paradox of Belief Revision precisely as challenging us to find a way of thinking of outright belief on which this does not collapse it into credence 1.

extent that the argument above makes trouble for this in-principle possibility, it would then threaten the less demanding conception just as much as the more demanding one.

Be that as it may, it is important to see that adopting even the most relaxed conception of rationality would only help up to a point. Let us suppose that satisfying the Appreciation Condition is *not* in any way necessary for rational belief revision. Even under this supposition, it remains plausible to think that, in the process of rationally revising our beliefs, we do *sometimes* appreciate the significance of contrary evidence, and carry out the revision on this basis.⁶ Call this kind of rational belief revision – a kind of rational belief revision that, by definition, satisfies the Appreciation Condition – *deliberative* belief revision. Given (1) and (2), deliberative belief revision is impossible for any subject like *S*. In other words, even if we waive the assumption that rational belief revision as such is subject to the Appreciation Condition, (1) and (2) will still force upon us (3’):

- (3’) Hence, if *S* believe that *p* (and considers the matter carefully enough) *S* cannot rationally *and deliberatively* revise her belief that *p*.

And (3’) seems no more palatable than (3). For even if not all rational belief revision is (or involves) deliberative belief revision, in the sense just defined, there is no question but that deliberative belief revision should be possible (both in general and for any subject like *S*). The idea that we *never* change our mind as a result of appreciating contrary evidence is intrinsically implausible. Indeed, one may argue that deliberative belief revision plays a vital role in at least some spheres of inquiry and debate, where we need to take special care in making up our minds and in what we offer to others to lead them to opinions.⁷ If this is right, the question whether belief revision that satisfies the Adequacy but not the Appreciation Condition qualifies as ‘rational’ need not detain us. Advocates and critics of the Appreciation Condition should all agree that the argument from (1) and (2) to either (3) or (3’) must be blocked. The question is only how.

2 The Too Early or Too Late Argument

Setting aside the assumption that rational belief revision must satisfy the Appreciation Condition, the Paradox of Belief Revision, as I formulated it in the last section, arises from only two premises.

Premise (1) rests on a principle of *closure* of rational belief under known entailment. Recall that *S* is a rational subject and her belief that *p* is a rational belief. Moreover, *S* knows that that (for any *p*) ‘*p*’ and ‘This piece of evidence supports the hypothesis that not-*p*’ entail ‘This piece of evidence is misleading’. Assuming that rational belief is closed under known entailment, it follows that *S* will believe, of any piece of evidence *E* that she takes to be evidence that not-*p*, that *E* is misleading evidence.

It hardly needs saying that closure principles which license this kind of inference are controversial. I will not attempt to defend these principles here, but I want to set aside responses to the paradox that involve denying them. This is partly because I think that

⁶ Cf. Wright’s (2014, 35) parallel remark about the Taking Condition.

⁷ I owe this point to Crispin Wright.

suitably qualified versions of the principles can be upheld in the face of standard criticisms (cf., for instance, Veber 2004, 561–562), and partly because the factors that are supposed to cause failures of the principles – such as context-shifts and risk-accumulation through premises conjunction – can be stipulated to be absent from the case we are considering (cf. Ye 2016, 564).

My conjecture, then, is that the paradox stands or falls with the truth of its second premise, which says that *S* cannot satisfy the Appreciation Condition with respect to any piece of evidence *E* that she takes to be evidence that not-*p* as long as *S* believes that *E* is misleading evidence. How must proponents of the paradox be thinking of the Appreciation Condition, if they are to use this premise in the way they do?

Suppose that Mary rationally revises her belief that John is in Paris after receiving a phone call in which John himself tells her that he is in Madrid: Mary goes from believing that John is in Paris to believing that he is in Madrid. Intuitively, what we expect from Mary is appreciation of three things. First, she should appreciate that John is telling her that he is in Madrid. Second, she should appreciate that John's telling her that he is in Madrid supports the hypothesis (or 'indicates') that he is in Madrid. Third, she should appreciate that, given what it indicates, John's testimony should bear (in the obvious way) on the question whether he is in Madrid or in Paris. Mary's change of mind, if it is to be rational, should result from her appreciation of (at least) each of these three elements. More in general:

Appreciation Condition: Rational belief revision should result from some appreciation of:

- (a) What the new evidence is ('It is the case that *p*')
 - (b) What the new evidence is evidence for ('That *p* indicates that *q*')
 - (c) Why the new evidence may require the relevant revision ('That *p* should bear on whether *q*')

In this formulation of the Appreciation Condition, the material in brackets states the proper content of the subject's appreciation – what occurs outside the brackets is just an informal, 'third-personal' gloss on what the subject is appreciating in appreciating that content. This means that the Appreciation Condition, as formulated above, does not require that the subject herself possess the concept of evidence, nor that they have true (or false) beliefs about their own beliefs.⁸

In principle, proponents of the Paradox of Belief Revision could support premise (2) by arguing that taking a certain piece of evidence to be misleading is incompatible with appreciating any of (a), (b) or (c) with respect to it. However, targeting (c) seems, by far, the most promising option. If one takes a certain piece of evidence to be misleading, one may still appreciate what evidence it is and what it is evidence for – indeed, it may be said that taking some evidence to be misleading presupposes such an appreciation. By contrast, it is at least initially plausible to think that someone who takes a certain piece of evidence to be misleading will not be able to appreciate the fact that such evidence should influence the answer to the question with respect to which he or she takes it to be misleading. If Mary

⁸ Importantly, this means that, for all I have said about it so far, the Appreciation Condition can be recognized to hold even by those who reject the Reflective Rationality Thesis.

thinks that John is lying (or is, for whatever reason, telling her something false), it seems that she will not be able to see any reason why she should change her mind on the basis of his testimony. This is just another way of saying that, if Mary takes John to be lying (or, for whatever reason, telling her something false), she will not be able to appreciate why John's testimony should influence the answer to the question whether John is in Madrid or in Paris. If one is interested in finding out the truth, it seems that one cannot, at the same time, regard a certain piece of evidence as evidence for a falsehood (i.e., as misleading evidence) and find such evidence relevant to one's inquiry.⁹

Note that this *prima facie* compelling thought can be sustained without taking on any controversial commitment on the nature of the act of appreciation. One familiar problem in spelling out what 'appreciation' means in this context is that we should not expect every instance (or even just most instances) of rational belief revision to involve explicit thoughts about one's evidence. Another problem is that subjects whom we want to regard as able of rationally revising their beliefs may not possess all the concepts that are needed to state the content of the act of appreciation. But note that, even if we take the Appreciation Condition to involve only an implicit and non-conceptual grasp of (c),¹⁰ it remains at least equally plausible to think that someone who regards a certain piece of evidence as misleading will not be able to appreciate (c) with respect to it. This means that proponents of the Paradox of Belief Revision need not enter the debate about how to interpret the word 'appreciate' – they can credibly say that their paradox arises on any proposed precisification of it.

A defence of premise (2) also need not assume any especially controversial account of the nature of belief. In particular, it need not assume an account of the nature belief on which one believes that *p* only if one is absolutely certain that *p*. If you believe that *p*, you take it for a fact that *p*, and if you take it for a fact that *p*, you ought to act as if *p* – whether or not you are absolutely *certain* that *p* seems irrelevant. For example, if you believe that the bridge is unstable, you ought to act as if the bridge was unstable even if you are not certain that it is. But acting as if a certain piece of evidence was misleading with respect to a certain question does not seem to be compatible with appreciating the fact that such evidence should influence the answer to that question. For insofar as one acts as if a certain piece of evidence was misleading with respect to a certain question, one is acting as if that question already had a settled answer – precisely the opposite answer than the one suggested by that piece of evidence. Thus, believing that one's new evidence is misleading seems enough to prevent one from appreciating (c) even if belief falls short of absolute certainty (as seems plausible).

⁹ It may be objected that, to the extent that one takes a certain piece of evidence *E* to be misleading evidence that not-*p*, one will *ipso facto* regard *E* as relevant to the question whether *p*. My point, however, is that one will not regard *E* as relevant to *one's inquiry* about whether *p* – or (to formulate the point without assuming that rational agents have thoughts about themselves and their inquiries) that one will not appreciate the fact that *E* should bear on whether *p*. Plausibly, appreciating the fact that *E* should bear on whether *p* implies being disposed to rely on *E* (possibly, in conjunction with other evidence) in one's reasoning about whether *p*. The objection misfires because one can regard *E* as relevant to the question whether *p* and yet fail to be disposed to rely on *E* in one's reasoning about whether *p*. The proponent of the Paradox of Belief Revision will suggest that this is exactly what happens when one regards *E* as misleading evidence that not-*p*.

¹⁰ This would be in analogy with recent attempts to spell out the Taking Condition (see footnote 3 above) in terms of intuitions, intellectual seemings, states of consciousness and the like. See, for instance, Chudnoff (2014), Hlobil (2019) and Marcus (forthcoming).

Perhaps, however, it is possible to attack (2) from a different angle. So far, we have been given reason to think that S cannot appreciate the significance of E as long as she regards E as misleading. But why think that the appreciation should occur at a time when S *still* regards E in that way? Here is an alternative model of rational belief revision. As Mary receives John's testimony (call it T), two things happen at once. On the one hand, T makes Mary's belief that John is in Paris irrational by providing support to the alternative hypothesis that John is in Madrid. On the other hand, as soon as Mary's belief that John is in Paris becomes irrational, her belief that T is misleading becomes irrational, too. Now, rational subjects are precisely those who (among other things) give up their beliefs when these become irrational. Since Mary is a rational subject, upon receiving John's testimony, she gives up both the belief that John is in Paris and the belief that his testimony is misleading. And this puts her in a position to appreciate the significance of John's testimony to the question whether John is in Madrid or in Paris.

The model suggests that premise (2) can be denied. It is not true that taking any contrary evidence to be misleading makes it impossible for S to satisfy the Appreciation Condition. The point is only that, in order to satisfy the Appreciation Condition with respect to any piece of evidence E , S must *first* give up any belief she may have that E is misleading. Appreciation can start as soon prejudice ends.

This response gives pause, but, ultimately, I think it is unlikely to convince proponents of the paradox – or indeed, anyone who finds any plausibility in the idea that rational belief revision must (or can sometimes) satisfy the Appreciation Condition. Remember that, according to the Appreciation Condition, rational belief revision must result from an appreciation of (a), (b) and (c). If the appreciation only starts when the revision has already taken place, there is no clear sense in which the latter can be said to 'result' from the former. If Mary begins to appreciate the significance of John's testimony only *after* changing her mind about John's being in Paris, it is not by appreciating the significance of his testimony that she changed her mind. We can perhaps make sense of what happened to Mary's beliefs in terms of some 'objective' relations of rational support holding between John's testimony and the hypothesis that he is in Madrid – but if we deny Mary any appreciation of the significance of that testimony till after the change, we cannot see her as displaying the kind of sensitivity to those relations that we should expect from a rational agent. At most, we will see her as behaving exactly *as if* she had such sensitivity.

Reflection on this point suggest a more stringent argument in support of premise (2). Let Q be some piece of evidence that S takes to be misleading evidence that not- p . S can appreciate the significance of Q to the question whether p either *before* or *after* losing her belief that Q is misleading. Before is *too early*: as long as S takes Q to be misleading, she cannot appreciate its significance. And after is *too late*: if S no longer takes Q to be misleading, she must have revised her belief that p already. Since there are no other possibilities, we should conclude that, insofar as S takes Q to be misleading, she cannot appreciate the significance of Q . And since the reasoning applies to any piece of evidence that S takes to be misleading, it follows that:

- (2) If S believes, of any piece of evidence E that she takes to be evidence that not- p , that E is misleading evidence, S cannot satisfy the Appreciation Condition (with respect to any such piece of evidence)

Let us call this the *Too Early or Too Late Argument*. The moral of this section is that, if we want to solve the Paradox of Belief Revision without calling the Appreciation Condition into question, we must find a way of resisting this argument. And that means that the choice we are now confronted with is between two options: shall we say that ‘before is not too early’ or that ‘after is not too late’?

3 Not Too Late or Not Too Early?

To make some progress towards answering our question, I suggest we make a short detour and look more closely at similarities and differences between the Paradox of Belief Revision and Kripke’s Dogmatism Paradox. This will allow us, first, to get clearer on what is truly distinctive about the Paradox of Belief Revision and, second, to identify the general shape that any satisfactory solution to either paradox should take.

Though the Dogmatism Paradox was originally formulated and subsequently published by Kripke (2011), its ‘canonical’ and best-known presentation is due to Harman:

If I know that h is true, I know that any evidence against h is evidence against something that is true; so I know that such evidence is misleading. But I should disregard evidence that I know is misleading. So, once I know that h is true, I am in a position to disregard any future evidence that seems to tell against h (Harman 1973, 148–9)

To make the parallel with the Paradox of Belief Revision most perspicuous, it is useful to turn Harman’s reasoning into a three-step argument:

- (I) If S knows that h is true, then (insofar as S considers the matter carefully enough) S knows that any evidence that not- h is misleading
- (II) If S knows that any evidence that not- h is misleading, S can disregard any future evidence that not- h
- (III) Hence, if S knows that h (and consider the matter carefully enough) S can disregard any future evidence that not- h

Here premise (I) rests on a principle of epistemic closure under known entailment: since S knows that it is false that not- h and that evidence supporting a false hypothesis qualifies as misleading, S knows that any evidence that not- h is misleading. Premise (II) can be defended by appeal to the principle that if you know a certain piece of evidence to be misleading you are justified to disregard it.¹¹ And what follows from (I) and (II) seems paradoxical because one would have thought that knowing something should not

¹¹ Harman is not explicit about what he means by ‘disregarding’. Lasonen-Aarnio reviews some interpretative options and argues that not lowering one’s credence in h in the light of E is a necessary condition for disregarding some evidence E that not- h (Lasonen-Aarnio 2014, 421–242). If one does not want to invoke credences, one can perhaps say that disregarding some evidence E that not- h (as it bears on the question whether h) is a matter of failing to be disposed to rely on E (possibly, in conjunction with other evidence) in any piece of reasoning about whether h . For my part, I shall assume only that disregarding some evidence E that not- h implies failing to appreciate the fact that E should bear on the question whether h (see footnote 9 above).

authorize the knower to disregard contrary evidence. Knowledge should not degenerate into dogmatism.

Let us focus on premise (II). Two subtle differences with the corresponding premise of the Paradox of Belief Revision merit attention. First, while both premises are conditional in form, the consequent of (II) affirms that *S can* do a certain thing ('...*S can* disregard...') whereas the consequent of (2) affirms that *S cannot* do another ('...*S cannot* satisfy...').¹² Dialectically speaking, this makes things more difficult for the proponent of the Dogmatism Paradox. In principle, the proponent of the Paradox of Belief Revision can limit herself to defending the aporetic conclusion that *S can neither appreciate nor* disregard any evidence that she takes to go against what she believes. By contrast, the proponent of the Dogmatism Paradox, insofar as she endorses (II), is committed to a positive view of what rational agents are justified to do with evidence that they know to be misleading – a view that (as we shall see in the next section) at least some of their opponents may criticize as implausibly permissive. Second, note that premise (II) of the Dogmatism Paradox involves a shift from *S's present* knowledge that any evidence that not-*h* is misleading to the possibility of disregarding any *future* evidence that not-*h*. The shift naturally invites another kind of response to the paradox, appealing to the defeasibility of knowledge: if we agree that the acquisition of new evidence may cause *S* to lose her knowledge, *S's* knowing that *h at one time* cannot be assumed to be the excuse for (or rational basis of) her disregarding evidence that not-*h at another time* – knowledge may have been lost in the meantime.¹³ Premise (2) of the Paradox of Belief Revision does not involve any analogous shift. And even if it did, the proponent of the paradox could plausibly defend its legitimacy using the Too Early or Too Late Argument.

This brings me to what is perhaps the most significant difference between the two paradoxes. One major reason why the defeasibilist reply seems so natural in the case of the Dogmatism Paradox has to do with the fact that the latter is concerned specifically with *knowledge*. We are (or can easily become) accustomed to the idea that new evidence may destroy someone's knowledge even if that person does not appreciate the significance of that evidence to the question at hand. This is because we are accustomed to the idea that, quite generally, knowledge can be destroyed by factors that lie *outside* the scope of the subject's appreciation. Most obviously, knowledge requires the obtaining of whatever state of affairs it is knowledge of, so it may be lost when the relevant state of affairs ceases to obtain. One way in which I may cease to know that there is milk in the fridge is by you deciding to remove the milk from the fridge. Perhaps new evidence may 'rob' me of my knowledge in the same way that your decision to remove the milk from the fridge does. But things seem to be different when, instead of knowledge, we focus our attention on rational belief, and try to make sense of the idea that we can change our beliefs in a rational way. If rational belief revision is subject to the Appreciation Condition, rationally revising one's beliefs cannot be a matter of being 'robbed' of one's beliefs through the acquisition of evidence of which one does *not* appreciate the significance. This is why responses

¹² See also other formulations of premise (II), according to which the subject is 'licensed' (Vahid 2012, 47) or 'has the right' (Harman 1973, 148) or 'is in a position' (Ye 2016, 565) to disregard contrary evidence.

¹³ The defeasibility strategy is advocated by Conee (2001). For discussion, see Lasonen-Aarnio (2014) and Ye (2016).

to the Paradox of Belief Revision appealing to *ex post* appreciation appear to be problematic: the acquisition of new evidence may result in some of one's beliefs becoming 'objectively' irrational, but we don't have any basis to speak of rational belief revision until this fact is, as it were, brought within the subject's view. And that is what the subject's belief that the evidence in question is misleading seems to make impossible.

There is, however, an important lesson to learn from recent discussions of the defeasibilist reply to the Dogmatism Paradox. Lasonen-Aarnio (2014) has convincingly argued that – whether or not it can form the basis of a satisfactory reply to Harman's original formulation of the paradox – the idea that new evidence can destroy the subject's knowledge by defeating the old evidence on which such knowledge rests cannot solve all problems in this area. To see why, suppose that our subject *S* knows that *p* and acquires some very weak evidence *E* that not-*p*. Specifically, suppose that *E* is so weak that, after acquiring it, *S* still knows that *p*. Still knowing that *p*, *S* knows that *E* is misleading evidence. And, assuming that one is justified to disregard evidence that one knows to be misleading, this means that *S* is still justified to disregard *E*. Now, Lasonen-Aarnio points out that, to the extent that we find (III) implausible, we should not be comfortable with this conclusion. Disregarding contrary evidence seems an unacceptable form of dogmatism. If we have the intuition that knowledge should not degenerate into dogmatism, we should be worried about the fact that, as long as it is *not* destroyed, knowledge can continue to justify its owner to disregard contrary evidence. So – Lasonen-Aarnio concludes – if we are to do justice to our anti-dogmatist intuitions, we cannot limit ourselves to embracing a defeasibilist picture on which knowledge can sometimes be destroyed by contrary evidence. Rather, we should take issue with the very idea that knowledge entitles one to disregard such evidence. In other words, we should deny:

Rational Disregard: As long as *S* knows that *E* is misleading evidence that not-*h*, it is rational for *S* to disregard *E* as it bears on whether *h*.¹⁴

A similar lesson applies to the case of rational belief revision. Suppose that our subject *S* believes that *p* and acquires some very weak evidence *V* that not-*p*. Specifically, suppose that *V* is so weak that, after acquiring it, it is *still* rational for *S* to believe that *p*. Still believing that *p*, *S* continues to believe that *V* is misleading evidence. And assuming that believing that a certain piece of evidence is misleading poses an obstacle to the possibility of appreciating the fact that such evidence should bear on the question at hand, this means that *S* will not be able to appreciate that *V* should bear on the question whether *p*. Now, even though we stipulated *V* to be very weak evidence that not-*p*, this conclusion seems problematic. *V* may not require *S* to change her mind, but this is no excuse for *S* to fail to appreciate that *V* should bear on the question whether *p*. Such appreciation – one might think – should be obligatory whether or not the evidence is strong enough to defeat the rationality of *S*'s prejudice

¹⁴ This is a slightly simplified version of a principle Lasonen-Aarnio calls 'Entitlement' (2014, 419).

against it. And if we want to vindicate the possibility of complying with this obligation, we should take issue with the idea that taking a certain piece of evidence to be misleading prevents the subject from appreciating its significance. In other words, we should deny:

No Early Appreciation: As long as S believes that E is misleading evidence that not- p , S cannot appreciate the fact that E should bear on the question whether p .

Remember that the last section ended with a question: in responding to the Too Early or Too Late Argument, shall we say that ‘before is not too early’ or that ‘after is not too late’? We have now uncovered a powerful (additional) reason to think that the answer should be that ‘before is not too early’. We cannot plausibly think that the kind of appreciation involved in rationally revising one’s own belief that p takes place *after* that belief has been given up. After is too late if we want the appreciation to be part of the process that results in the revision. But, crucially, it is also too late if we want the appreciation to take place in *every* case where the subject acquires contrary evidence, and not just in those cases where the contrary evidence is strong enough to make it rational for the subject to give up the belief. The right solution to the Paradox of Belief Revision should involve denying No Early Appreciation.

4 Belief Revision and Self-Reflection

Let us reconsider Rational Disregard, the principle that a subject who knows that a certain piece of evidence is misleading is rationally entitled to disregard it. It has been said that, even if it may initially sound plausible, this principle can be criticized as ‘bad epistemic policy’:

Acceptance of [Rational Disregard] overlooks the fact that we often take ourselves to know when we do not [...]. Cases will arise where we apply the principle to things we merely think we know, and by disregarding what could be corrective evidence, we force our heads deeper and deeper into the sand. [Rational Disregard] is bad epistemic policy. (Veber 2004, 567)

Veber’s complaint in this passage is essentially this: Rational Disregard says that it is rational for one to disregard evidence that one knows to be misleading, but if one were to use this principle as a policy or rule guiding one’s epistemic endeavours, one would often end up disregarding evidence that should not be disregarded – for we often take ourselves to know when we do not. Therefore, Rational Disregard should be rejected.

The complaint is not entirely fair, for – as Lasonen-Arnio points out – “the truth of [Rational Disregard] does not entail in any straightforward way that a subject ought to employ [a maxim like: ‘If you know that a piece of evidence is misleading, ignore it’] as a policy or rule guiding her belief-revision” (2014, 430). We can see Rational Disregard as offering a third-personal normative *evaluation*, not a first-personal normative *directive* – let alone one that subjects may be described as ‘following’ even when they disregard a certain piece of evidence only because they *think* they know it to

be misleading.¹⁵ But now consider the following variant of Rational Disregard, obtained by replacing ‘knows’ with ‘believes’:

*Rational Disregard**: As long as *S* believes that *E* is misleading evidence that not-*h*, it is rational for *S* to disregard *E* as it bears on whether *h*

Here a complaint along the lines of Veber’s seems perfectly apt and can plausibly be made without interpreting the principle as offering a ‘followable’ policy, rule or directive. When one disregards evidence that one *believes* to be misleading, one may be doing something irrational. Since our beliefs are often false, there are plenty of cases where we regard what could be corrective evidence as misleading. By disregarding such evidence, we would force our heads deeper and deeper in the sand. There is nothing rational about *that*. So – whatever the merits of Rational Disregard – Rational Disregard* should certainly be rejected.

The question is where the falsity of Rational Disregard* leaves us with respect to our goal, which is to undermine the plausibility of No Early Appreciation. If Rational Disregard* is false, it is not rational for *S* to disregard evidence that she believes to be misleading. But this does not yet show that it is possible for *S* not to disregard such evidence. In general, it is not obvious whether and how one can go from ‘It is not rational to φ ’ (or, for that matter, ‘It is rational not to φ ’) to ‘It is possible for a rational agent not to φ ’. And in this particular case, the transition is especially problematic: we can see perfectly well that *S* *should not* disregard the evidence, but – in order to have a satisfactory response to the Paradox of Belief Revision – we need a concrete account of how it can *make sense*, from *S*’s own perspective, not to disregard it – indeed, we need a concrete account of how *S* can bring herself to appreciate the significance of the evidence.

It is at this juncture that self-reflection can be seen to play an important (and, perhaps, irreplaceable) role. The suggestion I want to explore in the remainder of this section is that, insofar as a rational agent is aware of her own beliefs as beliefs she has, she will be able to ‘see’ a certain kind of risk involved in disregarding evidence that she takes to be misleading, and that ‘seeing’ that risk will allow her to appreciate why such evidence should in any case bear on the question at hand. Thus, the assumption that rational agents are systematically aware of their own beliefs as beliefs they have may put us in a position to deny No Early Appreciation and, thereby, see a way out of the Paradox of Belief Revision.

To begin to see the connection between the ability to form true second-order beliefs about one’s mental states and the ability to appreciate the significance of one’s evidence, it is useful to reflect on a type of reasoning that (though perhaps not frequent in everyday life) is philosophically familiar and relatively well-understood. Consider René, a rational subject endowed with the ability to form true second-order beliefs about his mental states. One day, a sceptic challenges René to produce evidence that he is not an envatted brain whose experiences are produced by the stimulations of a powerful computer, and not by the kind of ‘external world’ that we ordinarily take ourselves to inhabit. Initially, René is tempted to rule out the truth of the sceptical scenario by resorting to the simple fact that he has hands. But then the following line of thought occurs to him:

¹⁵ For the distinction between third-personal normative evaluations and first-personal normative directives, see Steinberger (2019).

“If the scenario described by the sceptic were true, I would *not* have hands, and yet it would *seem* to me as if I did. For if I were a brain in a vat, the computer stimulating my brain would produce in me exactly the same kind of experiences to which I owe my conviction that I have hands. Therefore, a certain *risk* is involved in letting the fact that I have hands bear on the question whether the sceptical scenario is true or not. It would be nice if I could avoid that risk and rest my response to the sceptic on a more solid basis...”

This is not the place to discuss whether it is possible for René (or anyone else in René’s position) to find a ‘more solid basis’ on which to refute external world scepticism. What interests me in this example is the notion of ‘risk’ it involves. While I do not pretend to have a fully worked-out account of this notion, I think that even a partial and intuitive grasp of it encourages two considerations which may prove helpful in connection with our discussion of No Early Appreciation.

First, in whatever sense it is ‘risky’ for René to respond to the sceptic by relying on the fact that he has hands, the relevant risk is one that René would not be able to ‘see’ if he did not have the ability to form true second-order beliefs about his mental states – specifically, beliefs about his own perceptual experiences. Note that it would not suffice for René to have the general notion of an illusory perceptual experience – for mere possession of that notion (or application of that notion to someone else) would not put René in a position to see the fact that he has hands under the dubious light that the sceptical scenario is meant to shed upon it. René needs to be aware that the fact that he has hands is also the content of a perceptual experience he is having if he is to be able to see that content as something that, under the circumstances described by the sceptic, would still *seem* to him to be fact without being one.

Second, insofar as he can see a ‘risk’ in his first response to the sceptic, René may be rationally motivated to avoid that risk. And, equally important, he can be so motivated even if he retains his belief (and, perhaps, knowledge of the fact) that he has hands. It may be that, further down the line, René’s epistemic perspective will change. He may end up a sceptic and revise his belief that he has hands. But, initially, the most immediate effect of his seeing the risk will be to prompt him to seek a different (‘more solid’) basis on which to rest his rejection of the sceptical scenario. And that is just to say that, by seeing the risk, René can come to appreciate that certain evidence, despite being perfectly good, should *not* bear on the question whether the sceptical scenario is true.

My proposal is that, in self-reflective agents, any case of rational belief revision involves a dynamic that is parallel and symmetric to the one we observe in René. As in René’s case, a certain risk is revealed to the agent thanks to his or her ability to form true second-order beliefs about his or her mental states. But, this time, what the agent comes to appreciate by way of perceiving the risk is that certain evidence, despite being (believed by him or her to be) misleading, *should* bear on the question at hand.

Another example will make the parallel vivid. Consider again Mary and her belief that John is in Paris. Let us assume that, just like René, Mary is a rational subject endowed with the ability to form true second-order beliefs about her mental states. Specifically, let us assume that Mary is endowed with self-reflection and, therefore, can form the true second-order belief that she believes that John is in Paris. When John calls Mary and tells her that he is in Madrid, Mary’s initial reaction is to dismiss John’s

testimony as irrelevant, based on what she takes to be a fact, namely that John is in Paris, not in Madrid. But then self-reflection puts Mary in a position to reason as follows:

“If the scenario described by John were true, John would not be in Paris, and yet it would *seem* to me as if he was. For I *believe* that John is in Paris and, when we believe something, what we believe seems to us to be a fact whether or not it is one. Therefore, a certain *risk* is involved in disregarding John’s testimony based solely on a fact that, if his testimony was true, wouldn’t be a fact at all. If I want to avoid this risk, I should allow John’s testimony to bear on the question whether he is in Paris or in Madrid...”

The two points I made above, in connection with René’s perception of the ‘risk’ involved in his first response to the sceptic, applies here as well. First, Mary would not be able to see the risk involved in her initial dismissal of John’s testimony if she were not a self-reflective agent. It will not suffice for her to have the general notion of a false belief – for mere possession of that notion (or application of that notion to someone else) will not put her in a position to see what she takes to be a fact (namely, that John is in Paris) as something that, in the scenario depicted by John, would *seem* to her to be fact without being one. Second, we can see Mary’s perception of the risk as affecting her attitude towards the evidence she has at her disposal. Maybe, further down the line, Mary will give up her belief that John is in Paris. But even as she clings to this belief and continues to regard contrary evidence as misleading, she can come to appreciate that (absent other, independent reasons to disregard it) such evidence *should* bear on the question whether John is in Madrid or in Paris.

The general picture of belief revision emerging from this example looks as follows:

- (i) Thanks to self-reflection, an agent who believes that p can become aware of the fact that she believes that p
- (ii) In being aware of the fact that she believes that p , the agent is aware that, if her belief that p was wrong, it would not be the case that p but it would still *seem* to her as if p
- (iii) Therefore, when the agent acquires a piece of evidence E that she takes to be evidence that not- p , she can see a certain risk involved in disregarding E based on (what she takes to be) the fact that p
- (iv) Seeing this risk and being rationally motivated to avoid it, the agent comes to appreciate that (absent other, independent reasons to disregard it) E *should* bear on the question whether p , even if she continues to take E to be misleading evidence.

I admit that this picture puts a heavy burden on the notion of ‘risk’, and that I have not said enough to put this notion on a firmer footing. One might be tempted to do that by linking risk with epistemic modality – for example, it might be suggested that (for any agent x) there is a risk that p if, and only if, (relative to x) it might be the case that p . But note that, on this view, the agent’s ‘seeing the risk’ would be a matter of her realizing that she ‘might’ be wrong in regarding the relevant piece of evidence as misleading. And it is unclear how the agent could realize *that* while, at the same time, retaining her belief that the piece of evidence in question is misleading. When one believes that p , one has made up one’s mind in such a way that one no longer regards its being the case

that not- p (and therefore, its being the case that one *wrongly* believes that p) as a live possibility.¹⁶ We must, therefore, understand (iii) and (iv) differently – not in terms of the subject's being open to the possibility of being wrong, but simply in terms of the subject's awareness of the fact that, if she were wrong, she would not be able to tell (at least, not based solely on her current evidential resources).

It may be complained that this kind of 'risk' (if it is one) cannot rationally motivate the agent to change her appreciation of the evidence – for the only risks that can do that are those representing possibilities that the agent has not ruled out. But this complaint begs the question against the present proposal. The point of René's example is that one can take sceptical arguments seriously – and be motivated by them to look for a 'solid basis' on which to refute them – while remaining convinced of the fact that one has hands.¹⁷ If we are to solve the Paradox of Belief Revision by denying No Early Appreciation, it is exactly this combination of attitudes that we need to make sense of. My tentative suggestion is that we can do so if we help ourselves to the assumption that rational agents are systematically aware of their own beliefs as beliefs they have.

5 Conclusions

The solution to the Paradox of Belief Revision outlined in the last section may not be the only possible one. Those who remain unconvinced by it may see this paper as an invitation to find an alternative way out. Three relatively uncontroversial points can nevertheless be extracted from the above discussion.

First, to the extent that there is some viable notion of belief revision that is subject to the Appreciation Condition (be it *rational* belief revision or *deliberative* belief revision, to use the terminology I introduced at the end of § 1) that notion will give rise to a paradox that (as discussed in § 3) has some traits in common with the Dogmatism Paradox, but also several distinctive features – chief among them that defeasibilism does not even begin to offer a solution to it.

Second, any satisfactory response to the Paradox of Belief Revision should involve the denial of No Early Appreciation, the principle, as long as S believes that E is misleading evidence that not- p , S cannot appreciate the fact that E should bear on the question whether p . In other words, solving the paradox requires showing that – strange as this may seem – a rational agent *can* appreciate the need to let a certain piece of evidence influence the answer to a certain question even if he or she regards that piece of evidence as misleading with respect to that question.

Third, if satisfying the Appreciation Condition is necessary for rational belief revision, and if the solution to the paradox outlined in the last section (or a fully worked out version of that solution) is the only possible one, then the Reflective Rationality Thesis holds true: rational belief revision requires an ability to form true second-order beliefs about one's own beliefs. The extent to which this line of thought, supporting the Reflective Rationality

¹⁶ This seems especially plausible if believing that p is, roughly, treating the proposition that p as if one knew it (cf. Williamson 2001, 46–47).

¹⁷ Cf. Beebe's (2010) argument that a sceptical scenario can raise an epistemological challenge (for an agent x) even if that scenario is not epistemically possible (for x).

Thesis, also supports a ‘constitutivist’ view of the relationship between rationality and self-knowledge is an issue that I shall leave for future investigation.¹⁸

Funding Open Access funding provided by Université de Genève.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Beebe, J. 2010. Constraints on sceptical hypotheses. *The Philosophical Quarterly* 60 (240): 449–470.
- Bilgrami, A. 1998. Self-knowledge and resentment. In *Knowing Our Own Minds*, ed. C. Wright, B. Smith, and C. Macdonald, 207–241. Oxford: Clarendon press.
- Boghossian, P. 2014. What is inference? *Philosophical Studies* 169 (1): 1–18.
- Burge, T. 1996. Our entitlement to self-knowledge. *Proceedings of the Aristotelian Society* 96 (1): 91–116.
- Chudnoff, E. 2014. The rational roles of intuitions. In *Intuitions*, ed. A.R. Booth and D.P. Rowbottom, 9–35. Oxford: Oxford University Press.
- Clarke, R. 2013. Belief is credence one (in context). *Philosophers' Imprint* 13 (11): 1–18.
- Conee, E. 2001. Heeding misleading evidence. *Philosophical Studies* 103 (2): 99–120.
- Davidson, D. 1963. Actions, reasons, and causes. *Journal of Philosophy* 60 (23): 685–700.
- Firth, R. 1978. Are epistemic concepts reducible to ethical concepts? In *Values and morals*, ed. A. Goldman and J. Kim, 215–229. Dordrecht: D. Reidel.
- Gallois, A. 1996. *The world without, the mind within: An essay on first-person authority*. Cambridge: CUP.
- Harman, G. 1973. *Thought*. Princeton: Princeton University Press.
- Hlobil, U. 2019. We cannot infer by accepting testimony. *Philosophical Studies* 176: 2589–2598.
- Kripke, S. (2011). ‘On two paradoxes of knowledge’. In his *Philosophical Troubles. Collected Papers Vol I*. Oxford: OUP: 27–51.
- Lasonen-Aarnio, M. 2014. The dogmatism puzzle. *Australasian Journal of Philosophy* 92 (3): 417–432.
- Marcus, E. (forthcoming). “Inference as consciousness of necessity”. *Analytic Philosophy*.
- Moran, R. 2001. *Authority and estrangement: An essay on self-knowledge*. Princeton: Princeton University Press.
- Reed, B. 2010. Self-knowledge and rationality. *Philosophy and Phenomenological Research* 80 (1): 164–181.
- Shoemaker, S. 1988. On knowing one's own mind. *Philosophical Perspectives* 2: 183–209.
- Steinberger, F. 2019. Three ways in which logic might be normative. *Journal of Philosophy* 116 (1): 5–31.
- Vahid, H. 2012. The dogmatism paradox and the problem of misleading evidence. *Teorema: International Journal of Philosophy* 31 (1): 47–58.

¹⁸ Earlier versions of this paper were presented at the Fifth Philosophy of Language and Mind Network Conference in St Andrews (August 2019) and at the Knowledge Beyond Natural Science Final Conference in Stirling (December 2019). I would like to thank the participants in these events for their feedback. I am also grateful to Peter Pagin, Crispin Wright and two anonymous referees of this journal for their detailed comments on the first draft. I owe special thanks to Fiona Doherty and David Mathers for helping me to develop the first idea of a paradox involving belief revision. In turning that idea into a paper, I received support from the Knowledge Beyond Natural Science Project (funded by the John Templeton Foundation at the University of Stirling from March 2017 to November 2019) and by the Swiss National Science Foundation (SNSF), through my Ambizione project “The Privileged Present: from Phenomenology to Metaphysics” (PZ00P1_186148).

- Veber, M. 2004. What do you do with misleading evidence? *The Philosophical Quarterly* 54 (217): 557–569.
- Williamson, T. 2001. *Knowledge and its limits*. Oxford: Oxford University Press.
- Wright, C. 1987. On making up one's mind: Wittgenstein on intention. In *Logic, Philosophy of Science and Epistemology*, ed. P. Weingartner and G. Schurz, 391–404. Vienna: Hölder-Pichler-Temsky.
- Wright, C. 2014. Comment on Paul Boghossian, 'what is inference'. *Philosophical Studies* 169 (1): 27–37.
- Ye, R. 2016. Misleading evidence and the dogmatism puzzle. *Australasian Journal of Philosophy* 94 (3): 563–575.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.