Chapter 12

Reflective Equilibrium

Yuri Cath

H. Cappelen, T. Gendler, and J. Hawthorne (eds.) *Oxford Handbook of Philosophical Methodology*, Oxford University Press (2016). [Preprint, please cite the published version]

1 Introduction

The method of reflective equilibrium is a method for figuring out what to believe about some target domain of philosophical interest like, say, justice, morality, or knowledge. This method is most closely associated with John Rawls who introduced the term 'reflective equilibrium' into the philosophical lexicon in *A Theory of Justice* (1971), where he appealed to this method in arguing for his famous theory of justice as fairness. However, Rawls had already advocated essentially the same method (without calling it 'reflective equilibrium') in an earlier paper (Rawls 1951), and Goodman (1954) is widely identified (including by Rawls himself) as being the first clear advocate of this kind of method, albeit with respect to a different normative domain, namely, logic.

Due to Rawls' influence, there has been a great deal of discussion of reflective equilibrium (or 'RE') in moral philosophy, and many philosophers have suggested that this method is uniquely suited to theorizing about morality or normativity more generally. However, the method of RE has also been claimed to be a method that plays a central role in all areas of philosophical inquiry, including those that are concerned with non-normative subjects, like metaphysics. This idea that the method of RE is employed in all areas of philosophy has been endorsed not only by proponents of this method (see e.g. Lewis 1983, pp. x-xi), but also by critics who agree that this method is used in this way but who argue that this is a practice that needs to be abandoned or reformed (see e.g. Cummins 1998; Stich 1990; Weinberg, Nichols and Stich 2001).

As well as different views of the scope and status of this method, there are also different views of what the method is, to the point where it can be misleading to talk (as is standard) of *the* method of RE. Accordingly, my first aim in this chapter will be to expose some of the subtleties involved in interpreting this method (§§2-4). I will then go on to consider some of the main objections to RE (§5). In closing, I will make some remarks about how the method of RE relates to recent debates about the role of intuitions in philosophy (§6).

2 An Initial Sketch

The aim of this section is to offer an initial sketch of the method of RE which will serve as a useful basis for our discussion. Following Scanlon (2003) and others, the method of RE can be usefully conceptualized as involving three distinct stages:

Stage 1: In this stage one identifies a relevant set of one's initial beliefs (or judgments or intuitions) about the relevant domain. These initial beliefs are often characterized as concerning particular rather than general features of the relevant domain. For example, if the domain is justice then one's set of initial beliefs might include the belief that a particular action was just, or if the domain is logic it might include the belief that a particular token inference was invalid, or if the domain is knowledge it might include the belief that a particular subject possesses knowledge, and so on. However, proponents of the method of RE often allow that the set of initial beliefs can also include beliefs about more general features of the relevant domain, including abstract general principles (see e.g. Rawls 1974, p. 8).

Stage 2: In the second stage one tries to come up with an initial set of theoretical principles that would systematise or account for the initial beliefs identified at the first stage. Scanlon (2003 pp. 140-141) in describing Rawls version of the method of RE writes that one is trying to come up with principles such that "had one simply been trying to apply them rather than trying to decide what seemed to be the case as far as justice is concerned, one would have been led to this same set of [initial] judgments".[1]

Stage 3: There are likely to be conflicts between one's initial beliefs and one's initial principles that aim to account for those beliefs. Furthermore, there are also likely to be conflicts between the members of one's initial set of beliefs, and perhaps even between the members of one's initial set of theoretical principles. These conflicts lead to the need for a third stage in which one engages in a reflective process of moving back and forth between these two sets and eliminating, adding to, or revising the members of either set until one ends up with a final set of beliefs and principles which cohere with each other. This final state is called a state of reflective equilibrium.

One point worth clarifying is that what the theoretical principles are meant to account for, and be in equilibrium with, is not the psychological fact that one has certain initial beliefs. This point is often obscured because: (i) descriptions of this method typically just speak of 'accounting for our beliefs', or 'bringing our beliefs into equilibrium'; and (ii) such descriptions are ambiguous given that propositional attitude terms like 'belief' can pick out both the psychological state of believ*ing* and the proposition that is believ*ed* to be true (as is the case for 'intuition' and 'judgment'). However, on close inspection, it is almost always clear that such descriptions should be interpreted in a non-psychologistic way. That is, when proponents of RE talk of theories 'accounting for our initial beliefs' they should be

[1] Sometime the method of RE is described in ways that do not fit so well with the characterisation I have given of this second stage. For example, I said that in the second stage one tries to "come up with" theoretical

interpreted as making a claim equivalent to something like 'accounting for the assumed truth of the contents of our initial beliefs', as opposed to 'accounting for the fact that we have certain initial beliefs'.[2] Similarly, when proponents of the method of RE talk of 'our beliefs being in a state of equilibrium' such claims are best interpreted as being equivalent to something like 'the contents of our beliefs being in a state of equilibrium' or 'our beliefs being in a state of equilibrium in virtue of their contents being in a state of equilibrium'. Some proponents of the method of RE do explicitly clarify this interpretative point,[3] but often it is left merely implicit, which can lead to unnecessary confusion and to misplaced criticisms.[4]

With that clarification in mind, we can explain how this method is meant to be one of figuring out what to believe about the target domain. The idea is that the process of bringing the contents of one's beliefs and one's theoretical principles into a state of equilibrium is one that should be mirrored by corresponding changes in one's belief states, so that by the end of this process the contents of one's resulting beliefs about the relevant domain should be captured by the final coherent set of propositions that one reaches in stage 3.

[2] To see this point it is useful to consider examples of the kinds of conflicts between our initial beliefs or intuitions and our initial theories that are meant to be resolved by the method of RE. Consider a familiar case from epistemology, namely, the conflict between the justified true belief ('JTB') analysis of knowledge and our intuition that the Gettier subject has non-knowledgeable justified true belief ('NKJTB'). This is a paradigm example of the kind of conflict between 'intuition' and 'theory' that the method of RE is meant to address. But, as Williamson (2007, p. 245-246) points out, it makes no sense if we interpret this conflict as one between the JTB analysis and the psychological fact that we believe or intuit that the Gettier subject has NKJTB because the assumed truth of the JTB analysis is consistent with that psychological fact.

[3] For example, Sayre-McCord (1996) makes this point clear when he writes: "The relative coherence of a set of beliefs is a matter of whether, and to what degree, the set exhibits (what I will call) *evidential consistency, connectedness,* and *comprehensiveness*. … Each…is a property of a set of beliefs, if it is at all, only in virtue of the evidential relations that hold among the **contents** of the beliefs in the set." (p. 166 bold emphasis added)

[4] For example, see the arbitrariness objection discussed below in §5.2.

Importantly, proponents of this method usually add the qualification that this state of equilibrium is an ideal that we should strive towards but will perhaps never achieve. For this reason, the method of RE is best viewed as a method that one is meant to continuously return to and reapply, rather than as a method that one would apply once and then set aside.

3 Interpreting the Sketch

To help fill in our initial sketch it will be useful to now consider a series of questions about how to interpret it.

What is meant to recommend this method of forming beliefs about the target domain? Proponents of this method hold that, when applied correctly, it will lead to beliefs that enjoy some positive normative status, where the most common suggestion is that these beliefs will be *justified* (or, at least, that one will have a justification to so believe). And proponents of this method often go further and suggest that it is the *best* and perhaps even the *only* method by which we can form justified beliefs about the relevant domain. For example, Scanlon (2003, p.149) endorses both these claims with respect to morality and other (non-specified) subjects:

> [I]t seems to me that this method, properly understood, is in fact the best way of making up one's mind about moral matters and about many other subjects. Indeed, it is the only defensible method: apparent alternatives to it are illusory.

Why think that the beliefs formed by this method would be justified? The method of RE is standardly interpreted as relying on a coherentist theory of justification and, indeed, is often referred to as simply being a "coherence method". This coherentist interpretation of the method of RE is sometimes disputed for reasons that we will discuss below in §4.3. But for now it will suffice to point out how this standard interpretation, if correct, provides a straightforward answer to the justification question, namely, that any belief formed by this method will be justified simply in virtue of it being a member of a system of beliefs that cohere with each other. On this interpretation then the method of RE is minimally committed to the claim that a belief's being part of a coherent system of beliefs is a *sufficient* condition for it being a justified belief.[5] And when proponents of RE suggest that this method is the *only* way of reaching justified beliefs about some relevant domain they appear to commit themselves to the idea that it is a *necessary* condition of a belief's being justified (at least for beliefs about the relevant domain) that it be a part of a system of beliefs that is coherent to some degree.[6]

What exactly does it mean to say that the beliefs one reaches at step 3 'cohere' with each other or are in a state of 'equilibrium'? Proponents of RE often do not provide much in the way of detailed answers to this question. One can find more detailed answers in the coherentism literature but there is no consensus account of what coherence is, and it is widely acknowledged that existing accounts are inadequate in different ways.[7] However, for our purposes, it will suffice to note that common to almost all accounts of coherence (in both the coherentist and RE literatures) is the very general thought that increasing the coherence of

[5] Or, alternatively, instead of appealing to the idea that it is sufficient for one's belief being doxastically justified one might merely appeal to the weaker claim that it is a sufficient condition for having propositional justification to so believe.

[6] I say 'to some degree' in relation to the fact that, as noted earlier, the state of reflective equilibrium is usually thought as an ideal that we should aim for but may never reach. But proponents of the method of RE will obviously want to say that we can still justify our beliefs (at least to some degree) by making steps towards this ideal.

[7] For example, Bonjour (1985) offered one of the most prominent and detailed accounts of the nature of coherence, but even he saw his account as "a long way from being as definitive as desirable" (1985, p. 101).

one's belief system is (at least partly) a matter of minimizing conflicts between the contents of one's beliefs, and maximizing certain relations of support between those contents. These notions of conflict and support are then analysed in a variety of ways by appealing to some mixture of deductive, probabilistic, evidential, or explanatory, relations between the contents of one's beliefs.[8]

What constraints, if any, are placed on the initial set of beliefs one identifies at stage 1? Some prominent proponents of this method—like Goodman (1953) and Lewis (1983)—place very few, if any, constraints on this set of initial beliefs. On the other hand, Rawls placed very specific constraints on the judgments that are the initial inputs into his version of the method of RE. According to Rawls, we should begin with only our "considered judgments", where he uses this as a technical term for those judgments which satisfy a range of constraints aimed at eliminating "judgments [that] are likely to be erroneous or to be influenced by excessive attention to our own interests" (Rawls 1971, p. 42). Rawls version of the method of RE can be thought of as one on which there is an intermediate step between stages 1 and 2 where one checks one's initial set of gathered judgments and filters out any that do not meet his constraints. These include the constraints that these judgments should not be ones made when one is upset or frightened, or where one's self-interests could be impacted by what the answer is to the relevant question. Rawls also requires that we only include those judgments

[8] How exactly should one bring one's beliefs into a state where they cohere with each other? For example, suppose that one's initial set of moral beliefs includes the belief that it would be morally impermissible for a surgeon to save the lives of five of their patients by giving them the organs of one of their other patients against the wishes of that patient (Foot 1967, Thompson 1985). Furthermore, suppose that one's initial set of moral principles include some simple act-consequentialist principle that, if correct, would classify this action as being morally obligatory. How should one resolve this conflict? Should one reject the initial belief or the theoretical principle or both? Again, proponents of the method of RE do not say as much about this kind of issue as one might like. But one idea that is present in many statements of the method of RE is that decisions about how to resolve such decisions should be sensitive to the *strength* of one's initial beliefs, as well as the *power* of the theoretical principles. See DePaul (1998, p. 295) for a nice discussion of these ideas.

in which we are confident, and which will be held stably over time. Another restriction that is sometimes placed on the initial inputs identified at stage 1 is that they have to be *intuitive* judgments or beliefs, and sometimes these inputs are described as simply being intuitions. Rawls (1951, p. 183) endorses a restriction of this kind, but it is worth noting that that he has a very minimal sense of this restriction on which an intuitive judgment is simply one that is not "guided by a conscious application of principles so far as this may be evidenced by introspection."

4 Many Methods

We have already indicated how the method of RE might be developed in subtly different ways depending, for example, on what restrictions one places on the initial beliefs identified at stage 1, or how one conceives of the nature of coherence. The aim in this section is to identify three more significant divisions between different interpretations of this method.

4.1 Deliberative versus Descriptive

We have been assuming that the method of RE is a method for figuring out *what to believe* about some target domain. Following Scanlon (2003), we can call this *the deliberative interpretation* of this method. My interest in this chapter is just in the deliberative interpretation. But it is worth noting, as Scanlon discusses, that Rawls himself seems to move back-and-forth between this deliberative conception and what Scanlon labels *the descriptive interpretation* of this method. On the descriptive interpretation, the aim of the method of RE is to reveal one's implicit *conception* of the relevant domain. For example, the method of RE aims at revealing our conception of, say, morality, as opposed to morality itself.

There may seem to be a tension between these two interpretations of the method of RE. However, Scanlon argues that, for Rawls, the descriptive version of the method depends on the deliberative version. This is because the way to reveal one's implicit conception of justice is to first figure out what to believe about justice itself by way of using the deliberative version of RE as described in §2. Whether employing the deliberative version of RE would actually be a good method for revealing one's conception of justice is, to my mind, far from clear. But, for our purposes, it will suffice to have just distinguished the deliberative interpretation from the descriptive, if only to put the latter aside.

4.2 Narrow versus Wide

Our initial sketch of the method of RE was a description of what is called the method of *narrow reflective equilibrium* (NRE) as opposed to what is called the method of *wide reflective equilibrium* (WRE). This distinction was first explicitly labelled as such in Rawls (1974), although it is usually thought of as being implicitly present in *A Theory of Justice* (as Rawls himself suggests in his 2001, p. 31), and most proponents of the method of RE endorse the wide version of this method.

The method of RE described in §2 is 'narrow' in two relevant senses: (i) it only aims at bringing two sets of things into a state of equilibrium, namely, the set of one's initial beliefs and the set of theoretical principles which are meant to account for those beliefs; and (ii) one's initial beliefs and theoretical principles are both about the same target domain. In contrast to (i), on the method of WRE one is trying to reach a state of equilibrium which will hold between these two sets and a third set of things, namely, any of one's other beliefs or "background theories" (Daniels 1974, p. 8) which are thought to be of some relevance to assessing one's initial beliefs and principles.

These further beliefs may also be about the target domain. For example, when Rawls (1974) introduces the notion of WRE what he appears to have in mind is an equilibrium between not only one's considered moral judgments and one's initial moral theories which are meant to account for those judgments, but also one's beliefs about the range of alternative moral theories and the arguments which are meant to support those theories. However, in contrast to (ii), these further beliefs may also be about other domains altogether. For example, in discussions of the method of WRE in moral philosophy it is often suggested that our beliefs about psychology, the theory of meaning, and metaphysics, might all potentially be relevant to figuring out what to believe about morality and, hence, that our idealized aim in moral theorizing should be a state of WRE that includes any such further beliefs insofar as they are relevant.

4.3 Coherentist versus Foundationalist Interpretations

As mentioned in §3, the method of RE is standardly viewed as being a coherentist method. However, a number of philosophers have suggested that this method is actually best interpreted as being committed to some kind of foundationalism. To help frame this issue it will be useful to make a distinction between *strong* versus *moderate* foundationalism, following Bonjour (1985, pp. 26-30). Strong foundationalism is a view that endorses something like the following two claims: (i) there are basic beliefs and (ii) these basic beliefs have some further special epistemic properties that make it a secure foundation for one's non-basic beliefs; where a basic belief is (roughly) a belief that is non-inferentially justified and which can justify other beliefs, and the kind of further epistemic properties that have been historically ascribed to basic beliefs include those of being "*infallible*, *certain, indubitable*, or *incorrigible*" (Bonjour 1985, pp. 26-30). Moderate foundationalism, on the other hand, is a view on which there are basic beliefs but they do not possess these further epistemic properties.

Proponents of RE do appear to be committed to denying strong foundationalism, that is, at least with respect to those domains to which that they think this method is applicable to. For, as we have seen, it is an important feature of the method of RE that *any* of one's beliefs about the relevant domain can, in principle, be rightly overturned on the basis of applying this method. And it is hard to see how one could square that assumption with the idea that some of those beliefs are not possibly false, or that they can't be rationally doubted, etc.

However, one could obviously reject strong foundationalism in this way and still endorse moderate foundationalism, which suggests that the method of RE is at least consistent with foundationalism. Furthermore, a number of commentators—including Ebertz (1993), Holmgren (1989), McMahan (2000), and Pust (2000)—have all suggested that the method of RE is best interpreted as being committed to the idea that the initial beliefs which are the inputs into this method must already be justified to some degree.

Often such claims seem to be motivated by a version of a standard worry about coherentism, namely, that increasing the coherence of one's beliefs can only "amplify" any justification already possessed by one's beliefs, but by itself cannot confer justification on one's beliefs. And, in order to avoid worry, it is sometimes suggested that the method of RE is best interpreted as being committed to some form of epistemic or phenomenal conservatism such that merely believing, or intuiting, that p is, in the absence of defeaters, a sufficient condition for having some degree of justification for believing that *p* (see e.g. Pust 2000, Ch. 1).

Interestingly, similar issues arise in the coherentism literature where many self-proclaimed coherentists endorse some form of epistemic conservatism. For example, Lycan (1998) offers a well-known explanatory form of coherentism, and on his view conservatism is one of the main explanatory virtues which can increase the coherence of a belief system. Furthermore, as Poston (2012, p. 78) points out, many historically important defenders of coherentism have endorsed related positions: “Lycan’s coherentism falls in line with the explanatory coherentist accounts of Goodman, Quine, Sellars, and Rawls by including a commitment to conservatism, the thesis that the mere holding of a belief confers some epistemic justification on its content.”

One might naturally think that what this shows us is that many supposed coherentists are not really coherentists at all, being instead proponents of moderate foundationalism. But this would be a mistake. For example, on Lycan’s view while “a belief is justified by the bare fact of its seeming to be true” (2012, p. 9) it is only so justified to a tiny degree and, crucially, it cannot justify other beliefs without the support of “other beliefs of varying grades of theoreticity, indeed relative to the subject’s entire belief system”. In which case, Lycan’s view is inconsistent with moderate foundationalism, as that is a view which claims that there are some beliefs which are both non-inferentially justified and which can, by themselves, justify other beliefs.

Lycan’s view is, at best, a version of a view that Bonjour calls *weak foundationalism*, according to which while there are non-inferentially justified beliefs they “possess only a very low degree of epistemic justification on their own, a degree of justification insufficient by itself either to satisfy the adequate-justification condition for knowledge or to qualify them as acceptable justifying premises for further beliefs” (Bonjour 1985, p. 28). Bonjour assumes that the commitment of this view to non-inferential justification suffices for it to be a form of foundationalism. However, Poston (2012) provides a strong case against this assumption. And Bonjour himself thinks (1985, p. 29) that this view is perhaps best seen as being a kind of hybrid of foundationalism and coherentism.

For our purposes, the interest of Lycan's position is that it helps us to assess the implications of these two claims about the method of RE: (i) that the input beliefs into this method must already be justified to some degree; and (ii) that this justification is non-inferential being based directly on the mere fact of the subject's believing or intuiting as they do. What Lycan's view shows us is how one can accept these claims and still offer a version of the method of RE that has strong coherentist features. In particular, one might offer a version of the method of RE on which an initial belief automatically possesses some degree of non-inferential justification, but this degree of justification is tiny and this belief can only participate in the justification of other beliefs insofar as it is part of a system of beliefs that approximates a state of reflective equilibrium. A view that would have the significant virtue of being able to accommodate the strongly coherentist statements made by Rawls and Goodman[9], as well as the fact that many coherentists identify their views as being closely related to, or simply versions of, the method of RE including Lycan himself (1998, p. 212) and others like Elgin (1996, Ch. 4) and Sayre-McCord (1996).

5 Objections

The most common kinds of objections made to RE all claim, in different ways, that this method is too weak (Kelly and McGrath 2010). That is, these objections all contend that one could apply this method perfectly to, say, one's initial moral beliefs, and yet the final set of beliefs that one would end up with, or the way in which one formed those beliefs, would still be criticisable in a way that undermines the credentials of RE as being a good method for

[9] For example, consider the following passage from Rawls (1971, p. 579): "I have not proceeded then as if first principles, or conditions thereon, or definitions either, have special features that permit them a peculiar place in justifying a moral doctrine. They are central elements and devices of theory, but justification rests upon the entire conception and how it fits in with and organizes our considered judgments in reflective equilibrium. As we noted before, justification is a matter of the mutual support of many considerations, of everything fitting together into one coherent whole."

figuring out what to believe about morality (or knowledge, or logic, etc.). And, typically, these weakness objections identify the reliance of this method on our initial beliefs or intuitions as being the source of the relevant problems. My aim in what follows is to identify some of the main objections of this kind and to indicate how proponents of the method of RE might reply to them.

5.1 Objections from Conservatism

One objection of this kind—that featured in prominent early criticisms of Rawls by Brandt (1979), Hare (1973) and Singer (1974)—is that the method of RE is a disguised form of moral intuitionism and, as such, is open to criticism for being too conservative in the importance it places on our moral theories conforming with our pre-theoretical moral beliefs or intuitions. This conservatism objection is often supported by suggestions that these initial beliefs may stem from untrustworthy sources. For example, Singer (1974, p. 516) writes that "all the particular moral judgments we intuitively make are likely to be derived from discarded religious systems, from warped views of sex and bodily functions, or from customs necessary for the survival of the group in social and economic circumstances that now lie in the distant past".

In response to this conservatism objection, proponents of RE often point to the fact that on this method the contents of one's initial beliefs are not meant to serve as fixed 'data points', to which any acceptable theoretical principles must conform. Rather, wherever there is a conflict between 'intuition' and 'theory' there is always the possibility that the best response to that conflict will be to reject one's initial beliefs. Furthermore, proponents of RE also point out that on the method of WRE there will be the potential for "far more drastic *theory-based* revisions of moral judgments" (Daniels 1979, p. 266).

One way of thinking of this revisability response would be to think of it as saying that while it is always a cost to a theory if it clashes with one's initial intuitions, this cost can be outweighed by the benefits of adopting that theory. But proponents of the method of RE often have a more radical form of revisability in mind such that the proper set of considered judgments against which our moral theories are to be "checked" (Rawls 1971, p. 51) should only be thought of as the judgments one holds at the *end* of the RE process (see e.g. Daniels 1979 fn. 17; and Scanlon 2003, p. 149).[10] And on this interpretation of RE it may be *no* cost to a theory at all if it clashes with our *initial* moral judgments or beliefs.

5.2 Objections from Disagreement

Despite the above revisability points, there is no doubt that our initial beliefs play a significant role in the method of RE. The worry can linger then that there is something objectionable about this fact. One source of concerns about this feature of the method of RE is the thought that different people may have very different initial beliefs and, hence, might reach different equilibria when they apply this method. For example, this thought can lead to the worry that if RE really is our best method for forming beliefs about a given domain then we will forced into adopting an anti-realist view of that domain.

The issue of whether the moral views of different people would converge after some form of idealised moral inquiry has often been thought to have implication for moral realism. Rawls himself (1974, p. 9) suggests that if the views of different people would not converge after they applied the method of WRE then it follows that there are no objective moral truths. It might also seem very plausible that there would be no such convergence given that there can

[10] See also DePaul's (1987) related distinction between 'conservative' versus 'radical' interpretations of the method of RE.

be substantial cross-cultural differences in our initial moral beliefs (Brandt 1979, p. 22). In which case, one objection that might be made to RE is that it can't be the best method of deciding what to believe about morality because assuming that it is will lead us to some form of anti-realism.

Of course, one may not view this supposed consequence of RE as constituting an objection to this method if one is prepared to abandon moral realism.[11] One might also be sceptical that there is any plausible method of moral inquiry such that everyone who impeccably employed that method would converge on the same moral views. In which case, this supposed consequence could not constitute a unique problem for the method of RE. But, perhaps most importantly, it is simply hard to see how the existence of moral disagreement is meant to support the denial of moral realism (see Enoch 2009 for a detailed discussion of this issue).

There are other disagreement-based objections that might be made to the method of RE. For example, one might object that this method is unjustifiably *arbitrary* in the way it relies on the initial beliefs or intuitions of the person who happens to be employing it, as opposed to the (perhaps conflicting) beliefs of someone else. After all, it would seem to be both egocentric and ethnocentric to adopt a method of inquiry which enjoins one to treat one's own intuitions, or the intuitions of one's culture, as being a reliable guide to the truth, whilst not assigning the same status to the intuitions of other individuals or cultures. As Ichikawa (forthcoming) discusses, Stich (1998) and Weinberg et al. (2001) both seem to appeal to something like this arbitrariness objection when they criticise the method of RE. Stich (1998) supports this objection by appealing to the possibility of cross-cultural disagreement in our epistemic intuitions, and Weinberg et al. (2001) support it by providing experimental evidence that there are actual disagreements of this kind.

[11] Rawls (1974) suggests that the method of WRE does not presuppose the existence of objective moral truths. See Holmgren (1987) for critical discussion of this claim, and Rawls (1980) for related discussion regarding his constructivist approach to moral theory.

One concern with this arbitrariness objection[12] is that it relies on a misunderstanding of the role that initial beliefs or intuitions play in the method of RE. In particular, it seems to rely on something (roughly) like the following two ideas: (i) the initial inputs into this method are propositions of the form *I/we have the intuition that p* which are then treated as evidence for *p*; and (ii) this method does not assign the same evidential role to propositions of the form *they have the intuition that p*. And if we think of the method of RE in this way then it is easy to understand why someone might view it as being egocentric or ethnocentric.

However, as discussed in §2, the inputs into the method of RE should not be understood in this psychologistic way. On the method of RE one does not begin (say) epistemological inquiry with propositions about one's epistemological intuitions. Rather, one begins inquiry with the contents of one's epistemological beliefs or intuitions. The rough idea being that one can justifiably treat the content of one's belief or intuition that *p* as being provisionally correct directly in virtue of one's believing or intuiting that *p*. As opposed, that is, to treating *p* as true in virtue of it being the conclusion of some inference based on the premise that one believes or intuits that *p*. Once one makes note of this point, the arbitrariness worry appears to lose much of its bite, as it is surely not arbitrary to begin inquiry with the contents of one's own mental states. Indeed, what else could we do?[13]

[12] The presentation given here of this arbitrariness objection differs from that found in Stich (1998) and Weinberg et al. (2001), as their version focuses on a specific worry about the normative force of conclusions reached by the method of RE. However, I think the kind of worries I raise above will also apply to their version of this objection. It is also important to note that more recent experimentalist critiques of the use of intuitions in philosophy do not rely on this kind of arbitrariness objection (see e.g. Weinberg 2007). See Ichikawa (forthcoming) for an excellent discussion which supports both these points.

[13] More should be said about this issue but for reasons of space these brief remarks will have to suffice here. For related discussion, see Wedgwood (2010) including the following passage (p.239-240): "It does not seem possible for me currently to form a moral belief *directly* on the basis of *your* moral intuitions. At best, I can only directly base my current formation of a moral belief on my *beliefs* about your moral intuitions. On the other hand, it *is* possible for me currently to form a moral belief directly on the basis of *my own current* moral

5.3 Objections from Error

The possibility of people reaching conflicting views when they properly employ the method of RE—because of their different initial beliefs—points to another important concern with this method, namely, that there is no guarantee that it will lead us to the truth. For if a method leads one person to believe that *p*, and another to believe that *not-p*, then, obviously, that method does not always lead to the truth. But the mere fact that a method of inquiry may lead to error, even when it has been impeccably applied, is not a good objection to that method. After all, as Kelly and McGrath (2010, p. 326) point out, the same criticism could be applied to the scientific method, as even the scientific method will consistently lead us to falsehoods if we have the misfortune to be in a world where "the empirical evidence that we have to go on is consistently misleading or unrepresentative".

Perhaps the error objection can be reformulated so as to avoid Kelly and McGrath's (henceforth 'K&M') overgeneralisation worry. One might try to argue that the problem with RE is that, unlike the scientific method, it is unlikely to lead one to the truth even when it is employed in "normal" conditions. But even if that idea could be made both precise and persuasive it is not clear that it would constitute a good objection. This is because proponents of RE are typically willing to concede that the method is not truth-conducive, whilst denying that this undermines the credentials of the method as a means of acquiring justified beliefs.

intuitions. Moreover, it seems that we are disposed to be guided by our moral intuitions towards forming the corresponding moral beliefs: if I currently have a moral intuition, that moral intuition will immediately incline me to accept the corresponding moral belief (unless I have some special reason for doubting that intuition)."

5.4 Objections from Unreasonable Belief

K&M (2010) argue that the key problem with RE is not that it is too conservative, or that it might lead us to false beliefs, but rather that it might lead us to beliefs that are unreasonable for us to hold. And if this claim is true it would constitute a powerful objection, given that what is meant to recommend this method is that it is a way to acquire justified beliefs (which presumably cannot be beliefs that are unreasonable for one to hold).[14]

According to K&M, even if the initial inputs into the method of RE have to meet Rawls' constraints on "considered judgments" it will still be possible that someone could impeccably apply this method and yet end up with beliefs that are unreasonable for them to hold. To support this claim, K&M give the example of a subject whose initial considered moral judgments include the judgment that the following proposition is true that I will label 'KILL':

> (KILL) One is morally required to occasionally kill randomly.

As K&M (2010, p. 347) note "there is nothing *incoherent* about the possibility that someone could confidently and stably subscribe to this judgment, even if he or she is aware of all of the non-moral facts, does not stand to gain or lose depending on whether it is true or false, and so on." But then it seems possible that a subject like this could impeccably apply the method of RE and end up with a final set of moral beliefs that still includes this perverse belief that KILL is true. Let us call this hypothetical subject 'Bill'. K&M think it is clear that it would not be reasonable for Bill to hold this perverse belief and, hence, they conclude that impeccably applying the method of RE does not suffice for one to end up with reasonable beliefs.

[14] If this claim is true it would also seem to constitute an objection to the DePaul's (1993 and 1998) version of the method of RE on which this method is not a reliable means of acquiring justified beliefs but is a means of acquiring rational beliefs.

K&M think the existence of this problem is obscured by the fact that when a proponent of RE illustrates her method she typically proceeds ‘from the first person perspective, and speaks of (e.g.) “our” considered judgments; she thus selects one of her own considered judgments that she expects her readers to share’ (2010 p. 347). But K&M think that this choice of example raises the worry that the basis for our agreement that the content of this judgment would be a good place to begin inquiry may not be the fact that it is the object of a considered judgment, but rather that it is a proposition which we perceive to possess some positive normative status, like being rationally credible or known to be true. K&M suggest that to decide between these two possibilities we should conduct a certain kind of ‘experiment’:

> In order to test the claim that it is the fact that the judgment in question is a considered judgment which is doing the work in this context, it is important to consider cases from the third person perspective, in which the starting points of the person pursuing reflective equilibrium are (i) his considered judgments but (ii) *perverse* considered judgments, at least when judged by one’s own lights. (That is, judgments which, when judged by one’s own lights, are clear cases of nonknowledge, or propositions utterly lacking in rational credibility.) (2010. p. 347)

And K&M claim that when one performs this experiment (e.g. by considering someone like Bill), one finds that “the idea that the normatively appropriate starting point for a person consists of all and only her considered judgments increasingly loses its appeal” (p.347).

According to K&M, the moral here is that the method of RE is only defensible if it is reconceived so that the initial beliefs it relies on do have to possess some positive normative status. As an example, K&M consider the possibility of a proponent of RE who rejects (1) in favour of (2):

> (1) For any individual, the appropriate starting point from which to pursue wide reflective equilibrium is the class of judgments consisting of all and only her considered judgments.

> (2) For any individual, the appropriate starting point from which to pursue wide reflective equilibrium is the class of all and only those judgments that she is justified in holding at that time.

K&M suggest that a proponent of RE who replaces (1) with (2) can avoid the perversity problem. But K&M also close their paper by querying whether the resulting method really deserves the name ‘the method of reflective equilibrium’.

What should we make of this perversity objection, K&M’s proposed solution to it, and the supposedly negative consequences of adopting that solution? Starting with the objection, K&M are surely right that it is possible for a subject to impeccably apply the method of RE to their initial beliefs and nonetheless end up with beliefs with crazy contents. But could a proponent of the method of RE still resist K&M’s further claim that we should thereby reject the method of RE (as it is standardly interpreted)?

One point that is worth noting is that K&M do not provide any arguments in support of their claim that it would not be reasonable for Bill to believe that KILL is true even after applying the method of RE. Presumably, this is because they take this claim to be too obviously true to warrant any such arguments. But I think there are good reasons to be wary of this claim that, in a way, mirror K&M’s own concerns about our agreement that a considered judgment would be a reasonable place to begin inquiry, when we ourselves endorse that same judgement.

K&M’s worry (as I understand it) is that in such cases our agreement might stem from our perception that the content of this judgment has some positive normative status (e.g. being rationally credible or known to be true). In which case, our agreement might not reflect a belief that this proposition would be an appropriate place to begin moral inquiry for anyone who has the considered judgment that it is true. Rather, it might merely reflect a belief that this proposition would be an appropriate place to begin moral inquiry for people like ourselves, who are aware that this proposition has this positive normative status, or who stand in some relevant normative relation to that content (like knowing it to be true).

The roughly analogous worry for K&M concerns their assumption that of a perverse proposition like KILL could not be part of an appropriate *end point* for inquiry into the relevant domain. The worry is that any inclination we have to agree with this claim might merely stem from our perception that there are overwhelmingly good reasons for rejecting this proposition. In which case, our agreement might only manifest a belief that this proposition could not be part of an appropriate end point of inquiry for people like ourselves, who are aware that this proposition has this negative normative status, or who stand in some relevant normative relation to it (e.g. knowing it to be false). As opposed, that is, to reflecting a belief that such judgments could never form an appropriate end point for anyone who inquired into that domain, no matter what initial beliefs they started with.

To support this worry, it might be useful to consider two different ways of thinking of Bill. On the first way we think of Bill's initial set of beliefs as being as close to our own as they could possibly be after adding just this one perverse belief. If we think of Bill in this way, it seems likely that if he impeccably applies the method of RE then this perverse belief should be quickly eliminated. This is because KILL will fail to cohere with all manner of other particular and general moral propositions that Bill initially believes to be true. On the other hand, we might think of Bill as having many other perverse moral beliefs (or, alternatively, we might think of Bill as only having this one perverse belief but then imagine that his confidence in this belief is far higher than his confidence in any of his other beliefs). If we think of Bill in this way then it seems likely that he will still end up with perverse moral beliefs after he applies the method of RE. But note that Bill's initial doxastic situation is now quite radically different from our own. And, once we think of Bill as being this different from ourselves, then I think it becomes far less obvious that it would be unreasonable for him to hold the perverse (from our perspective) beliefs that he might end up with after impeccably applying the method of RE.

For these reasons I think it is far from clear that K&M's perversity objection succeeds.[15] But, for the sake of argument, let us assume that it does. Does K&M's proposed solution to this problem succeed, and are they right to claim that this solution has negative consequences for the method of RE?

Starting with the solution, consider K&M's specific suggestion of replacing (1) with (2). I think it is clear that this proposed solution will fail to block the perversity problem. For example, consider, again, those versions of the method of RE that endorse some form of epistemic or phenomenal conservatism. These are views on which the inputs into the method of RE are meant to be justified beliefs. But it a well-known consequence of both epistemic and phenomenal conservatism, that a subject could be justified in holding crazy beliefs. This is because it seems perfectly coherent that there could be subjects who have perverse seemings or spontaneous beliefs whilst also being unaware of any relevant defeaters. In which case, if either epistemic or phenomenal conservartism is true then a proponent of RE cannot solve the perversity problem by simply replacing (1) with (2).

K&M might reply that this is not a problem with their solution to the perversity objection but rather a problem with conservatism. But I think this issue will generalise much further, indeed to almost any view which endorses the standard assumption that doxastic justification is non-factive. For if justification does not entail truth then it is hard to see how we can rule out the possibility of a subject having justified beliefs with perverse contents like KILL. In

[15] See Lycan (2012, pp. 11-12) for a broadly related defence of his coherentist theory of justified inference against the objection that it will have to sometimes classify crazy beliefs as being justified.

which case, it seems that a version of RE which restricts its inputs to justified beliefs will still face the perversity problem.

Perhaps one might try to save K&M's solution to the perversity problem by either defending the controversial thesis that doxastic justification is factive, or by appealing to some other normative status which is unquestionably factive (like, say, knowledge). But if the supposed moral now of the perversity objection is that the method of RE is only plausible if it is restricted to factive attitudes of some kind, then it seems like this objection is threatening to just collapse back into the objection that the method of RE need not lead us to the truth. But, as we have seen, K&M themselves point out that this kind of objection is problematic and their objection is meant to be independent of the error objection.

Our reflections on the perversity objection suggest that K&M have not provided a compelling case for thinking that the initial inputs into the method of RE have to possess some positive normative status. But, as we have already seen, there are proponents of the method of RE who already accept that the inputs into this method have to possess[16] some (perhaps tiny) degree of non-inferential justification. If only for this reason then, it is worth considering K&M's closing remarks in which they suggest that any such version of the method of RE will face certain negative consequences.

K&M suggest that if one requires that the initial inputs into the method of RE have to be justified beliefs then they one thereby lose one of the main supposed virtues of this method, namely, that it allows one to avoid positing some mysterious faculty as the source of our justified beliefs about the relevant domain. This is because one will now need to tell some further story about how our initial beliefs come to be justified. In which case, K&M claim

[16] Or provide, if we think of these inputs as being intuitions which are then understood as some kind of non-doxastic seeming state (see e.g. Pust 2000 for the suggestion that the inputs into method of RE should be understood in this kind of way).

that “the need for a certain kind of traditional epistemological theorizing (with all of its attendant pressures towards postulating non-obvious normative mechanisms, and so on) seems to have re-emerged” (2010, p. 353). Furthermore, K&M suggest that the resulting method may not deserve to be called a version of the method of RE because it will now be “natural to think that the most interesting part of the story concerns not the pursuit of equilibrium itself, but rather what makes it the case that certain starting points are more reasonable than others, and how we manage to recognise or grasp such facts” (2010, p. 354).

Starting with the interest worry, it is hard to see why we should accept this suggestion. The assumption that bringing one’s moral beliefs into equilibrium is only part of the story about how one’s beliefs get to be justified does not give us any reason to think that whatever else is needed to complete that story will be of greater interest than the equilibrium part of that story. Suppose a proponent of the method of RE grants that the initial beliefs identified at stage 1 must already be justified to some degree, and by something other than their coherence relations to other beliefs. This claim is perfectly consistent with the idea that coherence still plays a very significant and interesting role in justifying our beliefs. For example, as discussed in §3.3, one might hold that this degree of justification is tiny, and that initial belief can participate in the justification of other beliefs if they are part of a system of beliefs that is in state of reflective equilibrium.

What about K&M’s further worry that the explanation of how our initial moral beliefs are justified will have to be a story that appeals to some mysterious faculty of intuition? Again, it is hard to see why we should accept this suggestion. Consider, once more, those versions of the method of RE which endorse some form of epistemic or phenomenal conservatism. Whatever one thinks of such principles it is by no means obvious that they commit one to some mysterious faculty of moral intuition. Indeed, proponents of these principles often claim that one of their virtues is that it they can help to explain how we can acquire certain kinds of justified beliefs without appealing to such mysterious entities.[17]

6. Intuitions and RE

A lot more could be said about objections to the method of RE. However, in closing, I want to briefly comment on a different issue, namely, the relationship between this method and the idea that intuitions play a central role in philosophical inquiry.

The "experimental philosophy" (or "X-Phi") movement has helped to generate a large literature on intuitions and their supposed role in philosophy. Notably, the very first X-Phi paper—Weinberg et al. (2001)—identifies the method of RE as the most familiar example of *Intuition Romanticism*, which the authors' name for the intuition-based strategy for forming beliefs that is the subject of their experimental critique. The method of RE is also identified as an example of a problematic intuition-based method of inquiry in earlier works that influenced the X-Phi movement (see e.g. Stich 1998 and Cummins 1998), and defenders of the role of intuitions in philosophy have also identified the method of RE as an example of the general position they are seeking to defend (see e.g. Pust 2000, Ch. 1).

The method of RE has been strongly linked then with this widespread assumption that intuitions play a central role in philosophical inquiry. But, interestingly, I think the standard characterisations of the supposed role of intuitions in philosophy often diverge in significant ways from the role that our initial beliefs are meant to play in the method of RE.[18] One

[17] See e.g. Lycan (1998, pp. 207-215) for related discussion on the relationship between conservatism and the justification of our moral beliefs.

[18] There are also differences in the characterisations offered of the nature of our initial beliefs or intuitions. For example, as mentioned in §2, Rawls (1951) requires that his considered judgments be intuitive. But, as was also noted earlier, Rawls has a very undemanding notion of an intuitive judgment according to which they are simply

example concerns the role of intuitive counterexamples in evaluating theories. In the intuitions literature it is widely assumed that a conflict with our pre-theoretical intuitions always counts against a theory, even if that cost can be ultimately outweighed by the benefits of adopting that theory (see e.g. Weatherson 2003, p. 8). On the other hand, as we saw in §5.1, proponents of the method of RE suggest that a conflict with our initial beliefs or intuitions need not count against a theory at all if the content of that intuition is not judged to be true at the end of the RE process.

Another example concerns the assumption that intuitions "serve as a kind of rock bottom in philosophical argumentation" such that "*Intuitive judgments justify, but need no justification*" Cappelen (2012, p. 112).[19] As Cappelen discusses (2012, p. 118-122), this idea is difficult to make precise and can be developed in a number of different ways. But I think it is fair to say that this foundational-type role for intuitions is importantly different from the role that our initial beliefs are meant to play in the method of RE. As we saw in §4.3, some philosophers argue that the method of RE is best interpreted as being committed to our initial beliefs possessing some degree of non-inferential justification. But, as we also saw in §4.3, even if we grant this point there are good reasons to think that the method of RE should still be interpreted as holding that our initial beliefs can only participate in the justification of other beliefs insofar as they are part of a belief system that has been brought into a state of RE. This is not a view then on which our initial beliefs can justify without themselves needing justification.

Obviously, a lot more should be said about these two examples and the relationship between the method of RE and standard characterizations of the role of intuitions in philosophy. But I

judgments that are not consciously based on the application of theoretical principles. In which case, Rawls is not committed to the ideas—often found in the intuition literature—that intuitive judgments have a special phenomenology, or that they are based solely on our conceptual competences.

[19] Cappelen (2012) himself rejects this idea as well as the more general one that intuitions play a central role in philosophical inquiry.

think these brief considerations suggest that this relationship may be more complex and interesting than it is usually assumed to be.

References

BonJour, L. (1985). *The Structure of Empirical Knowledge*. Cambridge, MA: Harvard University Press.

Brandt, R. (1979). *A Theory of the Good and the Right*. Oxford: Oxford University Press.

Cappelen, H. (2012). *Philosophy Without Intuitions*. Oxford: Oxford University Press.

Cummins, R. (1998). "Reflection on Reflective Equilibrium". In DePaul and Ramsey (eds.) *Rethinking Intuition*. Lanham, MD: Rowman and Littlefield: 113–127.

Daniels, N. (1979). "Wide Reflective Equilibrium and Theory Acceptance in Ethics". *Journal of Philosophy* 76 (5): 256–282.

DePaul, M. (1987). "Two Conceptions of Coherence Methods in Ethics". *Mind* 96: 463–481.

DePaul, M. (1993). *Balance and Refinement: Beyond Coherence Methods of Moral Inquiry*. London: Routledge.

DePaul, M. (1998). "Why Bother With Reflective Equilibrium?" In Michael DePaul and William Ramsey (eds.) *Rethinking Intuition*. Lanham, MD: Rowman and Littlfield: 293–309.

Ebertz, Roger P. (1993). "Is Reflective Equilibrium a Coherentist Model?" *Canadian Journal of Philosophy* 23 (2): 193–214.

Elgin, C. (1996). *Considered Judgment*. Princeton, New Jersey: Princeton University Press.

Enoch, D. (2009). "How is Moral Disagreement a Problem for Realism?" *Journal of Ethics* 13 (1): 15–50.

Foot, P. (1967). "The Problem of Abortion and the Doctrine of Double Effect". *Oxford Review* 5:5–15.

Goodman, N. (1954). "The New Riddle of Induction". In his *Fact, Fiction, and Forecast*. Cambridge MA: Harvard University Press: 59–83.

Hare, R. M. (1973). "Rawls' Theory of Justice". *Philosophical Quarterly* 23: 241–252

Holmgren, M. (1987). "Wide Reflective Equilibrium and Objective Moral Truth". *Metaphilosophy* 18 (2): 108–124.

Holmgren, M. (1989). "The Wide and Narrow of Reflective Equilibrium". *Canadian Journal of Philosophy* 19 (1): 43–60.

Ichikawa, J. (2014) "Who Needs Intuitions? Two Experimentalist Critiques". In Anthony Booth and Darrell Rowbottom (eds.), *Intuitions*, Oxford: Oxford University Press.

Kelly, T. and McGrath, S. (2010) “Is Reflective Equilibrium Enough?” *Philosophical Perspectives* 24 (1):325–359.

Lewis, D. (1983). *Philosophical Papers, Volume I.* Oxford: Oxford University Press.

Lycan, W. G. (1988). *Judgment and Justification*. New York: Cambridge University Press.

Lycan, W. G. (2012). “Explanationist Rebuttals (Coherentism Defended Again)”. *Southern Journal of Philosophy* 50 (1): 5–20.

McMahan, J. (2000). “Moral Intuition”. In Hugh LaFollete (ed.) *The Blackwell Guide to Ethical Theory*. Chichester: Blackwell: 92–110.

Poston, T. (2012). “Basic Reasons and First Philosophy: A Coherentist View of Reasons”. *Southern Journal of Philosophy* 50 (1): 75–93.

Pust, J. (2000). *Intuitions as Evidence*. New York: Routledge.

Rawls, J. (1951). “Outline of a Decision Procedure for Ethics”. *Philosophical Review* 60: 2:177–97. Reprinted in Rawls (1999): 1–19.

Rawls, J. (1971). *A Theory of Justice*, 2nd edition 1999. Cambridge MA: Harvard University Press.

Rawls, J. (1974). ‘The Independence of Moral Theory’. *Proceedings and Addresses of the American Philosophical Association* 47: 5–22. Reprinted in Rawls (1999): 286–302. Page references are to the reprinted version.

Rawls, J. (1980). “Kantian Constructivism in Moral Theory”. *Journal of Philosophy* 77: 515–572. Reprinted in Rawls (1999): 303–358. Page references are to the reprinted version.

Rawls, J. (1999). *Collected Papers*, Sam Freeman, (ed.). Cambridge MA: Harvard University Press.

Rawls, J. (2001). *Justice as Fairness*. Cambridge MA: Harvard University Press.

Sayre-McCord, G. (1996). “Coherentist Epistemology and Moral Theory”. In Sinnott-Armstrong, W. and Timmons, M. (eds.), *Moral Knowledge? New Readings in Moral Epistemology*. New York: Oxford University Press: 137–189.

Scanlon, T. M. (2003). “Rawls on Justification”. In Samuel Freeman (ed.), *The Cambridge Companion to Rawls*. New York: Cambridge University Press: 139–167.

Singer, P. (1974). “Sidgwick and Reflective Equilibrium”. *The Monist* 58 (3):490–517.

Stich, S. (1990). *The Fragmentation of Reason*. Cambridge MA: MIT Press.

Thomson, J. J. (1976). “Killing, Letting Die, and the Trolley Problem”. *The Monist* 59 (2): 204–217.

Weatherson, B. (2003). “What Good Are Counterexamples?” *Philosophical Studies* 115 (1): 1–31.

Wedgwood, R. (2010). “The Moral Evil Demons”. In Richard Feldman and Ted Warfield (eds.), *Disagreement*. Oxford: Oxford University Press: 216–246.

Weinberg, J. (2007). “How to Challenge Intuitions Empirically Without Risking Skepticism”. *Midwest Studies in Philosophy* 31 (1): 318–343.

Weinberg, J., Nichols, S., and Stich, S. (2001). “Normativity and Epistemic Intuitions”. *Philosophical Topics*, 29 (1–2): 429–460.

Willimason, T. (2007). *The Philosophy of Philosophy*. Oxford: Blackwell Publishing.