Reflective Equilibrium

Antti Kauppinen & Jaakko Hirvelä

For the *Oxford Handbook of Normative Ethics*

Revised draft, August 5, 2022

Introduction

No one ever begins ethical inquiry without already having many ethical convictions. We start from a position in which we already believe that, say, happiness is good and torment bad, and that people shouldn't lie whenever it benefits them. According to what has come to be called the *method of reflective equilibrium* (RE for short), we should take such convictions seriously in moral inquiry, unless we have specific reason to doubt them, and try to formulate general principles that would explain their truth. We should also consider independent arguments for the candidate principles, and be willing to modify or give up our considered judgments to make them fit our principles, as well as vice versa. If, as a result, our moral principles and considered judgments are in harmony, and also fit with the rest of what we believe, and we know these things, we have reached a state of reflective equilibrium.

This initial sketch is very rough, since there are many ways to formulate RE – indeed, in Section 2, we will introduce a schema that allows the construction of 256 different variants. One important point of divergence concerns the very aims of the method. Nevertheless, the most common thing to say is that beliefs that are in reflective equilibrium are in some sense *justified*. Accordingly, we will mostly focus on this issue, particularly on whether RE is a source of *epistemic* justification. From this perspective, there are two key defining features of the method. The *negative* epistemic claim of RE is that no moral beliefs are (fully) justified unless they cohere with one's other moral and non-moral beliefs. This makes RE in one sense *non-foundationalist*. The *positive* epistemic claim of RE is that the favored kind of coherence among the relevant elements suffices

for being epistemically justified in believing moral propositions. This makes RE *coherentist* and *non-skeptical*.

It is not unusual for people to say that there are no real alternatives to the method of reflective equilibrium in ethics (e.g. Scanlon 2002; DePaul 2006, 616–618). But rejecting either defining epistemic claim yields two clear alternatives. First, if we reject RE's *negative* claim, we'll take some moral beliefs to be justified independently of their fit with other moral beliefs. Some hold that certain moral beliefs are justified *non-inferentially* (without needing support from other beliefs), either because they are *self-evident* (Sidgwick 1907, Ross 1930, Audi 2003) or because we're capable of *moral perception* (Audi 2013). Others are *rationalists* who hold that moral beliefs can be justified by some kind of transcendental argument from the very possibility of rational agency (Kant 1785/2000; Korsgaard 1996; de Maagt 2017).

Second, we may reject RE's *positive* claim, denying that coherence suffices for justification. Foundationalists who reject RE's negative claim typically also reject the positive claim, holding that all justification derives from intuition (etc.). It's worth noting that if we *accept* the negative non-foundationalist claim and simultaneously reject the positive coherentist one, we risk moral scepticism. After all, in that case we're saying that *neither* intuition (etc.) nor coherence suffices for justification. But as we'll see, there is room for a more subtle view, which holds that both coherence and non-coherentist partial justification are needed for full justification. So even if coherence doesn't suffice for justification, it may play a more modest but crucial justificatory role.

We'll address these issues as follows. We will begin with a brief overview of the idea of reflective equilibrium in the work of John Rawls and others and what we will call the "epistemic turn" in its use. Next, we'll work through a fairly detailed example of the method, and give a systematic overview of the different possible ways of specifying it. We'll then examine various criticisms and the most promising responses. In the final section, we argue that instead of an account of *alethic* epistemic justification (the kind of justification that bears on the truth of a

proposition), RE may be best understood as a method of *moral inquiry* (and possibly a kind of *dialectical justification*), since engaging in the kind of reflection involved in the method of RE is plausibly the best feasible way to achieve *moral understanding* and the ability to justify one's convictions *to* others.

1. The Idea of Reflective Equilibrium

The *locus classicus* for the idea of reflective equilibrium in ethics is John Rawls's brief methodological section in his epoch-making *A Theory of Justice*. It is inspired in part by Nelson Goodman's earlier work on logic (Rawls 1971, 20). Goodman had argued that the only way to justify deductive or inductive logical principles is in terms of their conformity with the particular inferences we're actually willing to make, and *vice versa*. As he acknowledged, this is blatantly circular – but according to him, virtuously so:

> A rule is amended if it yields an inference we are unwilling to accept; an inference is rejected if it violates a rule we are unwilling to amend. The process of justification is the delicate one of making mutual adjustments between rules and accepted inferences; and in the agreement achieved lies the only justification needed for either. (Goodman 1955, 67)

Departing somewhat from Goodman, Rawls (1971) doesn't directly try to balance moral principles and particular verdicts against each other. Instead, he famously introduces a device of representation he calls the *original position*. Its conditions are meant to guarantee that whatever principles of justice are agreed to by all rational agents occupying it will be fair. This is achieved by having the parties seek their own interest behind a 'veil of ignorance' that screens off facts that shouldn't influence what is due to a person, such as natural talents, social position, or specific ideas about what constitutes a good life. Rawls believes that principles adopted in the original position are consequently justifiable to all actual people regardless of their talents, position, or conception of the

good, as long as they're willing to treat everyone as free and equal. (This notion of dialectical moral justification is something we'll come back to at the end.)

Our concern here is not with the particular principles that Rawls believes would be agreed to, but with how they should be evaluated. He says that we test the conditions of the original position by considering whether the principles of justice they entail "match our considered convictions of justice or extend them in an acceptable way" (1971, 19). On some issues, we're quite certain about what justice requires – for example, that racial discrimination is unjust (Rawls 1971, 19) – and Rawls suggests that we can treat such convictions as 'provisional fixed points' that candidate principles should match.[1] On other issues, such as those concerning the just distribution of wealth, we're antecedently less confident, and should correspondingly be more willing to be guided by principles that would be chosen in the original position.

What if there is a mismatch between our considered convictions and the chosen principles? Rawls says:

> We can either modify the account of the initial situation or we can revise our existing judgments, for even the judgments we take provisionally as fixed points are liable to revision. By going back and forth, sometimes altering the conditions of the contractual circumstances, at others withdrawing our judgments and conforming them to principle, I assume that eventually we shall find a description of the initial situation that both expresses reasonable conditions and yields principles which match our considered judgments duly pruned and adjusted. This state of affairs I refer to as *reflective equilibrium*. (Rawls 1971, 20, emphasis ours)

[1] In his later work on political liberalism, Rawls qualifies the 'we' whose convictions are at issue in political justification by saying that the initial starting points are matters of overlapping consensus among reasonable comprehensive doctrines in a liberal society. See the end of Section 4 for more on this.

As Rawls continues, it is an *equilibrium* because the principles and particular judgments coincide after mutual adjustment, and it is *reflective* "since we know to what principles our judgments conform and the premises of their derivation" (ibid.). He explicitly conceives of this as a coherentist conception of justification:

> A conception of justice cannot be deduced from self-evident premises or conditions on principles; instead, its justification is a matter of the mutual support of many considerations, of everything fitting together into one coherent view. (1971, 21)

For Rawls, the theory-independent inputs are *considered* judgments or convictions in which "our moral capacities are most likely to be displayed without distortion" (1971, 47). So we set aside judgments we lack confidence in, that are ill-informed, or are made at the height of emotion, or are likely to be biased by self-interest (ibid.; Rawls 1951). When it comes to principles, they may be presented on their own or with supporting arguments. Rawls says that equilibrium is *wide* when it results from consideration of all plausible conceptions and their supporting grounds (1974, 8). This equilibrium is individualistic – when Rawls raises the issue of who the relevant 'we' are, he notes that for his purposes, only the views of "the reader and the author" count (1971, 50).

The notion of a wide reflective equilibrium was influentially recast as a *method* for theory choice in ethics by Norman Daniels (1979). Emphasizing the importance of explanatory coherence beyond balancing principles and considered judgments, Daniels conceives of the relevant arguments for principles as inferences from a *background theory*. Consider, for example, Rawls's assumptions about the nature of persons, which support his contractualist account of justification (Daniels 1979, 261). Among other things, Rawls assumes that persons are fundamentally separate from each other, so that benefits to one person don't morally compensate for harms to another. In contrast, on Derek Parfit's view of personal identity, interpersonal boundaries are 'metaphysically shallow' and not very important, so that benefits to one person *do* compensate for harms to another, much as later

benefits for me may compensate for present harms to me (Parfit 1984, 341). This conception supports utilitarian principles that endorse interpersonal compensation. While these background theories are philosophical, Daniels holds that depending on the question, empirical theories about psychology, society, economics, and so on, may also be relevant for moral views, and thus yield reasons for preferring one theory over another independently of match with considered judgments.

What is the point of seeking reflective equilibrium? While Rawls clearly talks about *justifying* a conception of justice, his methodological remarks suggested an alternative *descriptive* picture. He says that in moral theorizing, "we may bracket the problem of moral truth" and instead "investigate the substantive moral conceptions that people hold, or would hold, under suitably defined conditions" (1974, 7). It is in order to do this, he asserts, that we try to find principles that match people's judgments in reflective equilibrium (ibid.) These are striking claims. They suggest that for Rawls at this point, we seek reflective equilibrium to understand what people *are* really committed to, not what people *should* be committed to, and thus do psychology rather than ethics or epistemology (see Mikhail 2011 for development of this idea). It is therefore not surprising that when he later revised *A Theory of Justice*, he eliminated some (though not all) of the language suggestive of this purely descriptive claim.

Indeed, we think it is fair to say that most later writers have largely ignored Rawls's descriptive self-understanding, and treated reflective equilibrium as a method of *epistemic justification*. Along Daniels's lines, the textbook understanding is that the method of reflective equilibrium is a way to test and refine moral theories (e.g. Kagan 1998, 12–16). In this context, one often finds appeals to "our" intuitions or considered judgments. The other typical characterization is found in work in moral epistemology, where the aim of the method is said to be epistemic justification of an *individual's* beliefs. Geoffrey Sayre-McCord, for example, defends the coherentist view that "as one approaches a (wide) reflective equilibrium one thereby increases the extent to which the beliefs one holds are epistemically justified" (Sayre-McCord 1996, 143; cf.

DePaul 1987, Brink 1989, Tersman 1993). Indeed, many recent objections to RE only make sense insofar as it is understood as an account of epistemic justification, as Folke Tersman (2018) notes. Thus even if RE was introduced as a method for vaguely understood moral justification or for investigating our actual moral conceptions, there has been an *epistemic turn* in the way it is standardly understood.

2. Varieties of Reflective Equilibrium

We've seen that the idea of reflective equilibrium has been put to different uses, whose relations to each other are not obvious. In this section, we will systematically consider different ways of fleshing out the basic idea.

Let us start with some definitions. Let's tentatively say that for S's beliefs to be in *reflective equilibrium* is for her first-order ethical beliefs, beliefs in moral principles or theories, and beliefs in background theories to cohere with each other (this is the "equilibrium" part) and for S to know that this is the case (this is the "reflective" part). Such reflective equilibrium is in one sense 'wide', since it includes coherence with background beliefs. Since no one defends the sufficiency of a more narrow equilibrium, we'll only consider the wide version here. Second, *the method of reflective equilibrium* is a process of purposefully seeking to arrive at reflective equilibrium by way of filtering out unreliable first-order beliefs, considering as many initially credible principles or theories as possible as potential explainers of the truth of the filtered beliefs, drawing on background theories to construct and weigh arguments for and against the candidate principles or theories, and mutually adjusting the filtered beliefs and best-fitting, background-theory-supported principles to arrive at a coherent system that manifests general theoretical virtues like simplicity and fertility. It is plausible that we can only arrive at a full *reflective* equilibrium through some such process of sorting out our beliefs.

To get a firm grasp of what these definitions mean, let's work through an example of the method of reflective equilibrium as a tool of epistemic justification. Suppose that the following are representative examples of the contents of Oona's ethical beliefs:

(A) Hard work deserves a reward.

(B) We should tax people and corporations to provide a good life for everyone.

(C) It shouldn't matter to your chances in life who your parents happen to be.

(D) Everyone is worthy of respect, which means that we shouldn't interfere with their economic choices unless they choose to harm others.

(E) No one should have to suffer from poverty when there are enough resources for everyone.

(F) The current US economic system is unjust.

(These are relatively general convictions – Oona will no doubt have views about the justice of particular cases, but we lack the space to articulate such scenarios here.)

Now Oona wants to figure out what makes a society's distribution of wealth just. She listens to some podcasts on ethics, and narrows the contenders down to the following two views:

(P1) Distributive justice requires ensuring that resources are divided in a way that maximizes total well-being in the long run, which means taking from the rich and giving to the poor.

(P2) Distributive justice requires ensuring that everyone has fair and equal access to social advantage and that they get to keep what they deserve according to rules that maximize economic benefit to the least well-off.

Oona also has a number of background beliefs that are relevant to assessing the candidate principles (evidently, they are also conclusions of possibly complex arguments that we can't discuss here):

(BB1) Determinism is true and incompatible with moral responsibility.

(BB2) Calm and dispassionate reflection is likely to lead us toward moral truth, but the truths of metaphysics lie beyond human reason.

(BB3) If you take from the rich and give to the poor, the well-being gains of the poor are greater than the losses of the rich.

To decide between P1 and P2, Oona reflects on how they fit with A, B, C, D, E and F, as well as BB1, BB2, and BB3. First, though, she notes that she's not too confident about F – on reflection, she realizes she might only believe it because she envies the superrich or because she doesn't really understand how the economic system works, so she sets it aside. It evidently takes some work and additional assumptions to figure out what P1 and P2 entail for the truth of the remaining beliefs we've listed – for example, she may need to consider actual or hypothetical scenarios of, say, hard-working people being rewarded or not being rewarded. But when she does that work, she observes that neither P1 nor P2 is initially a perfect fit with all the considered judgments, though since both do seem to explain C and E, she can set them aside for the time being.

How do P1 and P2 compare with each other? In addition to C and E, the roughly utilitarian P1 would explain the truth of B (taxing the rich), in the light of BB3 (basically, decreasing marginal utility). But Oona realizes that it is likely to clash with A (deserving reward) and D (respect for rights), because total well-being may in some circumstances be maximized by redistributing things regardless of desert or rights. But then again, she also believes BB1 (absence of personal responsibility), which also clashes with A and plausibly D. So P1+B+C+E+BB1+BB3 form a nice, coherent system of beliefs, which she can enjoy simply by giving up A and D.

The broadly Rawlsian P2 has different strengths and weaknesses relative to what Oona believes. In addition to explaining C and E, it is a pretty good match with D (respect for rights), though it requires somewhat more extensive interference, if people's choices result in resource

inequality that undermines equality of opportunity. It also explains A (deserving reward), assuming that rewarding hard work benefits the worst-off, and isn't far from B (taxing the rich to provide a good life for everyone), although it demands ensuring only a good start in life for everyone, leaving it open that they end up poor through their own life choices. On further reflection, Oona finds she's willing to adjust B in this way – after all, why should hard-working people sacrifice their gains to support the hedonistic lifestyle of surfers? (cf. Van Parijs 1991) Even if she was initially confident in B, the very process of reflection can change her initial credences (DePaul 1987). But what about BB1 – if people are not free and responsible for their choices, isn't it better to give up A and D altogether rather than make minor adjustments? Here BB2 (the priority of ethics to metaphysics) comes into play. Oona is more confident of her ethical beliefs than her views on the metaphysics of free will, so when they clash, she's willing to modify the latter. What's more, when she encounters the further argument that people can be *entitled* to their holdings without *deserving* them (Nozick 1974, Rawls 2001), she realizes that the clash between her metaphysical and moral views need not be so deep.

So let us say that when Oona considers these and further arguments for the background beliefs and the theories' implications for further concrete cases, she finds that P2 is the best match for her considered convictions. It's not a perfect match, so she'll have to make some adjustments, highlighted in the following revised list:

(P2') Distributive justice requires ensuring that everyone has fair and equal access to social advantage and that they get to keep *what they're entitled to* according to rules that maximize economic benefit to the least well-off.

(A') Hard work *entitles one to a reward*.

(B') We should tax people and corporations to pay *for education and healthcare*.

(D') Everyone deserves respect, which means that we shouldn't interfere with their economic choices unless they choose to harm others *or result in so much economic inequality that it undermines fair equality of opportunity*.

(E') No one should have to suffer from poverty when there is plenty enough resources for everyone*, unless they suffer as a result of free choices*.

Her moral beliefs in the area of economic justice form a coherent whole that is unified through a fairly simple principle – though of course, to keep the exposition manageable, we've only considered a small fragment. It will be an important question, as we will explain, just what follows from such coherence.

Oona's method is just one possible way of filling out the details of RE. Indeed, while Oona's procedure is evidently inspired by the classic accounts discussed in the previous section, it differs in various ways from all of them. To get a more systematic grasp of the possible elements of the method of reflective equilibrium, let us introduce the following schema for generating variants:

Subject: *an individual* or *a group* (decision point A)

Inputs: 1) *considered judgments* or *intuitions* (B) about *particular cases* or *at any level of generality* (C), 2) *individual principles* or *moral theories* (roughly, sets of principles together with some rationale for them) (D), 3) *arguments based on background theories/beliefs* (wide RE) or *nothing in addition to judgments/intuitions and principles/theories* (narrow RE) (E)

Processing: filtering out inputs that are formed in ways that are unlikely to be truth-conducive in general *by one's own lights* or *as a matter of fact* (F)

Principle of adjustment: *bottom-up* (maximizing coherence among beliefs or credences in individual propositions) or *top-down* (plausibility of whole alternative systems) (G)

Source of justification: *coherence alone* or *coherence among independently credible inputs* (H)

Aim: *epistemic justification* of principles/*theory choice in ethics*, *moral justification* of principles or theories, *identifying people's actual moral convictions*, or *responsible inquiry* (I)

We've already introduced options at points A-E, and remaining ones will be discussed in the following. Briefly, when it comes to processing (F), what counts as a considered judgment may be determined either by what the subject takes to be distorting influences on judgment *or* by what in fact are such. Mutual adjustment among principles and judgments (G) may be guided by the subject's levels of confidence (or *credences*) in them and in relations among them (as in Oona's reasoning), or by the intuitive plausibility of whole packages, where the latter option gives many more degrees of freedom (cf. Brandt 1979, 18–21). We'll come back to H and I in Sections 3 and 4, respectively.

Different choices at the 8 first decision points of this schema yield 256 different possible variants of the method of reflective equilibrium (1024 if we count the aim as part of a conception of RE). We will return to the question of which is the best way of specifying the method in the final section, bearing in mind that it depends on one's theoretical purposes. But before that, we must look at the various criticisms that have been levelled against RE understood in one way or another.

3. Criticisms of Reflective Equilibrium

While the idea of reflective equilibrium has been enormously influential, it has also been the target of frequent criticisms. We'll begin with general issues with coherentism about epistemic justification and move on to issues specific to ethics.

Let's start with the question of what it *is* for someone to have epistemic justification for believing that *p* (see e.g. Alston 2006). On one kind of view, for you to have epistemic justification for believing that *p* is for you to be *epistemically blameless* or possibly *epistemically praiseworthy* or *responsible* for believing that *p*, on the basis of what justifies it (Bonjour 1985, 8). On another, for you to have epistemic justification for believing that *p* is for you to *meet your epistemic obligations* if you believe that *p* (Littlejohn 2012, 4), or for you to have an *epistemically permissible* belief (Pollock 1986, 125). These two can come apart, since you may have only an excuse but not justification for believing that *p*, and thus be blameless (and possibly in a sense praiseworthy in virtue of displaying rationality) in spite of failing to live up to epistemic norms (Williamson forthcoming, Littlejohn forthcoming). The distance between being blameless and conforming to epistemic norms will depend in part on one's view of what the relevant epistemic norms are. It is common to hold that justification has something to do with truth – other things being equal, we would expect justified beliefs to be more likely to be true than unjustified ones. As it is sometimes put, if a belief is justified, it is reasonable to hold it when one aims to believe truths and avoid falsehoods (Tersman 1993, 14). As we'll see next, this makes trouble for coherentism.

3.1 Does Coherence Justify Belief?

Suppose that we take the aim of RE, as is common these days, to be arriving at a coherent set of beliefs that is therefore epistemically justified. Perhaps individual beliefs in this set are justified when the whole system is more coherent with them than without them (Feldman 2003, 65). But what exactly is coherence, and does it suffice to justify?

Let's start with the former question. A plausible idea is that the coherence of a set of beliefs is a function of how their contents relate to each other. It seems that a coherent set of beliefs must be *logically consistent*, but no one thinks this suffices for justification – after all, any random beliefs whose truth values are independent meet this criterion. Some even deny that it is necessary for

justification (Hirvelä 2022). Instead, coherentists talk about *mutual support* among beliefs. But what is mutual support? For empirical beliefs, one prominent suggestion appeals to *probabilistic support*. According to C. I. Lewis (1946), a belief that *p* coheres with a set of beliefs just in case the conditional probability of *p* given the rest of what we believe is higher than its probability apart from the rest of what we believe, and vice versa. However, such views are ill-suited for current purposes, because basic moral truths are widely thought to be necessary truths, so that their conditional probability is the same as their unconditional probability.

Perhaps coherence should then be understood in terms of *inferential* and *explanatory relations* among propositions, as Laurence Bonjour (1985), among others, argues. That is, the contents of some beliefs can serve as premises for a cogent argument for others, or the truth of some beliefs jointly explains the truth of others. However, the latter, at least, requires a very different notion of explanation than we have in the case of empirical and contingent truths, insofar as in the moral case we're concerned with necessary truths that are trivially entailed by everything.

Does coherence, understood in inferential or explanatory terms, suffice to justify belief? The traditional objection is that it doesn't, because the contents of false beliefs can stand in all the same inferential or explanatory relations to each other as those of true ones, so that whether beliefs *p*, *q*, and *r* mutually support each other doesn't make it any more likely that they're true than if they don't. The most obvious coherentist reply is to endorse a form of constructivism, according to which moral truths are metaphysically determined by coherent sets of beliefs – for example, torture is wrong in virtue of the fact that belief that torture is wrong is part of a coherent set of moral beliefs (Street 2006). But this line of response is not very popular, because the wrongness of something like torture seems to be entirely independent of our beliefs about it, and because it entails the possibility of radical relativism (torture is wrong *for me*, given my beliefs, but possibly not *for you*).

Alternatively, coherentists can endorse a form of internalism about justification: if each of my beliefs is supported by my other beliefs (that is, other things I take to be true), then *from my perspective*, there is reason to think it's likely to be objectively true (Tersman 1993, 101). If I hold my beliefs because of recognizing such support relations, it is at least somewhat plausible that I'm epistemically blameless in believing as I do, even if my beliefs are in fact false, and thus in one sense justified. But this arguably gets believers off the hook too easily. Even if you as a matter of fact hold false beliefs, it is possible that you would have been *able* to form correct beliefs, had you been, say, more conscientious or more open-minded. And if your belief that one job candidate is better than another is based on a prejudice, but you don't believe you're prejudiced, the biased belief can cohere perfectly with what you take to be true (Lemos 2018, 378–9; cf. Srinivasan 2020). Thus, you might merit epistemic criticism even if your beliefs support each other. And of course, if justification is more closely linked to truth, this move is a non-starter in conjunction with objectivism about moral truth.

Perhaps coherence ought to be understood in terms of relations among our *attitudes* rather than their *contents*. Oona recognized that she wasn't very confident about F, and set it aside in order to maintain the ethical beliefs that she was more confident in. The way in which she updated her *credences*, i.e. her degrees of confidence, contributed to the coherence of her epistemic position. In epistemology, Bayesians hold that a subject's prior credence distribution must be probabilistically coherent and that the subject should update her credences by applying the Bayes theorem ("conditionalizing"). Within this framework we can understand inferential and explanatory connections in terms of strict inequality between unconditional and conditional credences. That is, S's evidence E supports *p* to a certain degree, if, and only if S's credence in *p* given E is higher than her unconditional credence in *p*.[2] Unfortunately, it can be shown that in the absence of some

[2] An immediate problem for this view is that coherent probability functions assign all necessary truths the prior probability 1, and hence nothing could explain a necessary truth. Bayesians can reply that what is actually necessarily true need not be necessarily true from one's perspective, and that the prior probability distribution is fixed by one's perspective. This view can make sense of our uncertainty of moral truths.

independent grounds to think that certain evidence is truth-conducive, coherence among one's degrees of belief does not yield justification in the Bayesian framework (cf. Olsson 2002).

These observations pose a dilemma for non-skeptical coherentists in ethics. Either they need to provide an account of epistemic support that entails that justification can emerge from a coherent set of individually unjustified beliefs, or they ought to accept that at least some beliefs have to have a positive epistemic standing that is independent of the subject's other beliefs. We think that the prospects of formulating plausible epistemic support-relations that would fit the bill are dim. Thus, as Richard Brandt argued, the starting points of reflection must be "initially credible … for some reason other than their coherence" (1979, 20). If coherentists endorse this option, they will effectively abandon coherentism in favour of *weak foundationalism*.

And indeed, many defenders of RE have recently defended weakly foundationalist versions of the view. Some draw on the epistemically conservative idea that either any belief we currently hold, or whatever seems to us to be true, is automatically defeasibly justified (Pust 2000). Ralph Wedgwood argues that some of our ethical intuitions are bound to be reliable in virtue of what it takes to possess normative concepts, and the role of RE is to weed out the errors that may infect them (2006, 81; see also DePaul 1986, Elgin 2014, 245, 267-8; Baumberger & Brun 2021, 7935). Mark van Roojen (2015), in turn, holds that RE is needed, because while we have *some* degree of non-inferential justification with respect to intuited propositions, it does not by itself suffice to license full belief in them in the absence of coherence with other similar propositions. However, an "only somewhat reliable intuitive judgment generating process can be part of an overall reliable process of reaching reflective equilibrium about a subject matter that is more reliable as a result of incorporating it" (van Roojen 2015, 155). As James van Cleve describes it, according to such weak foundationalism, coherence can *amplify* initial warrant, and may thus be required for full justification (2011, 338; cf. Haack 1993).

As long as the initial credibility of the propositions that serve as the starting points of reflection does not rule out rejecting or modifying them for coherence, weak foundationalism is entirely compatible with RE. Since this helps RE avoid the problem with coherentism, we take this to be part of the best formulation. The crucial question then concerns initial credibility. Fortunately, it does not seem all that implausible that we have at least some justification to believe that our considered judgments are true. After all, on many versions of RE, those judgments need to be formed in ways that are as a matter of fact (and not just in the subject's own opinion) not prone to lead us astray, and this fact alone might render those judgments *prima facie* justified. Indeed, when Rawls first sketched his view in 1951, he talked about the judgments of *competent judges* as the starting points of theory construction. If competence with a subject matter entails a degree of reliability, even (early) Rawls may have been a weak foundationalist. It is worth noting, however, that on this version of RE, justification isn't *transparent*: we might falsely believe that our starting points are sufficiently reliable, and thus lack justification in spite of having a coherent set of beliefs.

So far, we've focused on the insufficiency of mere coherence for justification. But it's worth emphasizing that it is plausibly also not *necessary* for justification. On anyone's view, we're justified in believing what we know. So if I know that it is wrong for a military unit to bomb civilians who neither pose nor are responsible for a threat to anyone, I'm justified in believing so. And it seems that I could know this sort of thing even if my moral beliefs are not coherent – say, I subscribe to some misguided moral theory. Generally speaking, if I can have any moral knowledge without a coherent moral belief system, coherence can't be necessary for justification.

3.2 Garbage In, Garbage Out

Even if we set aside general worries about coherence, we might think that there is something *especially* problematic about balancing candidate moral principles against independently held judgments or intuitions. The general form of this argument is the following:

Garbage In, Garbage Out

1. The outcome of RE depends crucially on the pretheoretical intuitions that serve as inputs to the process. (From definition of many variants of RE)
2. All (or a certain class) of our pretheoretical moral intuitions are influenced by many factors that have nothing to do with moral truth. (Garbage In)
3. If the outcome of a process that results in beliefs about a subject matter depends on factors that have nothing to do with the truth about the subject matter, the resulting beliefs are not epistemically justified.
4. So, beliefs that result from RE are not epistemically justified. (Garbage Out)

The third premise draws on the idea that it can't just be a matter of luck whether justified beliefs are true. Again, while most epistemologists think that there can be justified but false beliefs as well as true but unjustified ones, there must be some intimate connection between justification and truth. The crucial issue then becomes the truth of premise 2, Garbage In. In an early response to Rawls, Peter Singer defended it by asking the following rhetorical question:

> Why should we not rather make the opposite assumption, that all the particular moral judgments we intuitively make are likely to derive 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 which case, it would be best to forget all about our particular moral judgments, and start again from as near as we can get to self-evident moral axioms. (1974, 516)

(Note that Singer also here articulates a clear alternative to RE, a form of intuitionism that appeals to self-evidence.)

RE skeptics have pointed to two broad kinds of empirical evidence in support of Premise 2. One line of argument appeals especially to the influence of natural selection on the contents of our moral intuitions. Very briefly, it is plausibly adaptive in the evolutionary sense to have feelings or intuitions to the effect that pain is bad and murder is wrong, because such thoughts deter us (even in the face of temptation) from doing the kind of things that hurt our or our group's chances of survival and reproduction, at least in the long term. This promotes the spread of the genes that program for them (e.g. Joyce 2006). But thoughts that are fitness-enhancing need not be true, and especially in the moral case, adaptiveness and truth may come wide apart. If so, our moral intuitions are shaped by factors that have nothing to do with moral truth.

The second argument for Garbage In appeals to psychological findings. Many studies have found that people's moral intuitions depends on how the alternatives are described even if they're in fact identical (framing effects, Petrinovich & O'Neill 1996), whether they are feeling disgust or other emotions, even if they have nothing to do with the target of evaluation (Schnall et al. 2008), whether harm to others is caused up close and in person or at a distance (Greene et al. 2009), the order in which alternatives are presented (order effects, Schwitzgebel and Cushman 2012), and many other things. Since such influences have nothing to do with moral truth, Garbage In follows.

Defenders of RE have responded to this critique in various ways. One natural response is to emphasize the appeal to people's *considered judgments*, not unreflective intuitions (Greenspan 2015; cf. Kauppinen 2007, Brun 2014). Many of the studies mentioned clearly do not test for judgments that meet the Rawlsian criteria. But as critics reply, it's still implausible that filtering judgments would remove *all* of these influences, and indeed existing evidence suggests that prompting people to reflect before judging seems to make little difference (Paulo 2020; for a partial reply, see Kauppinen 2019).

It seems to us that the best response on behalf of RE has two prongs. The first begins with the observation that the critics of intuition must rely on *normative premises* when making the claim

that some influence on moral judgment has nothing to do with the truth about the subject matter. For example, one very plausibly adaptive moral judgment is that the survival of our family members is good, since it bolsters our willingness to help those who share our genes when the going gets tough. Whether evolution pushes our judgments in a morally irrelevant direction thus depends in part on whether it is true that the survival of our family members is good. As David Enoch points out, it very plausibly *is* true that survival is good – and pain is bad, and so on for many adaptive moral beliefs – so that at least in some important cases, evolution influences us at least roughly in the right direction (Enoch 2010, 428), so Garbage In is false. As RE rightly holds, we can't help making use of moral beliefs in assessing the epistemic significance of the causes of our moral beliefs (Tersman 2018, 4). While there is plenty more to say on this topic, it seems unlikely that evolutionary considerations by themselves warrant a wholesale skepticism about moral intuitions (Vavova 2021).

However, this line of response is less plausible when applied to the psychological findings. Factors like physical distance or framing of alternatives *are* very plausibly morally irrelevant. But does this show that RE is problematic? Defenders are sometimes tempted by a kind of gotcha-response: it seems that when critics reject certain intuitions as problematic, they are engaging precisely in the sort of reasoning that RE requires (Tersman 2008). After all, it is part of wide RE that we look at our considered judgments in the light of background theories. When the critics say that some judgments (like intuitions that suggest there is a difference between causing harm as a means and causing harm as a side effect) should not be considered, they do so because they conflict with other beliefs that we hold with more confidence (like the moral belief that the difference between causing harm up close and causing harm remotely is morally irrelevant, and the empirical belief that the means/side-effect intuitions result from this feature). So such arguments end up buttressing RE, because it gives a principled explanation of why we should not include the problematic beliefs in the process of figuring out what to believe. It doesn't follow that *all* of our

moral beliefs should be set aside – indeed, we must rely on *some* of them in judging the status of processes that shape our beliefs. It is only if there is an Archimedean point that doesn't require support from considered judgments that we can dispense with RE.

To be sure, this offers only limited support for RE, since it just shows that people who are aware of intuitively distorting influences on their beliefs can make use of RE to purify their list of considered judgments. Naïve subjects will still be licensed by RE to rely on judgments that in fact result, say, from disgust. Insofar as we think such judgments are epistemically unjustified, many variants of RE will still turn out to be too liberal in their starting points to provide epistemic justification.

3.3 Lack of Convergence

It is widely accepted that RE is a *path-dependent* process: where one ends up depends on initial starting points and choices made along the way (for example, regarding whether to adjust judgments or principles when they clash). Given that people start with different moral convictions and may make different choices, it is predictable that there are multiple reflective equilibria. Some such disagreements are likely to remain even if people are aware that others hold different views and try to reconcile their own views with them to the extent they find it reasonable to do so. As Thomas Kelly and Sarah McGrath put it,

> Different individuals might impeccably employ the method of (Rawlsian) reflective equilibrium and end up with substantially different moral views, even if they were exposed to all feasible moral conceptions and all reasonable arguments for those conceptions. (Kelly and McGrath 2010, 340)

Rawls (1974; 1985) thought that such divergence calls objective moral truth into question, which in part motivated his move to a political notion of justification (see below). However, lack of

convergence by itself does not imply the absence of epistemic justification or objective truth. As Kelly and McGrath (2010, 341) rightly point out, it is a hallmark of realism about a subject matter that no method is guaranteed to reach the truth about it. So this is not the problem with divergence. Kelly and McGrath argue instead that the problem is that faithfully following RE can lead to *unreasonable* beliefs, if one's initial considered judgments are bad enough (2010, 346). For example, someone could be in a reflective equilibrium while believing they should randomly kill people. For them, this shows that not all considered judgments merit consideration, so that RE is overinclusive in its starting points. But as Yuri Cath (2016) notes, for someone to arrive at badly wrong moral beliefs via RE, it doesn't suffice that they start with one or a few bad considered judgments (because they would end up discarded in the process), but must instead have a whole set of problematic beliefs to begin with, and most likely problematic background theories as well. In such a case, Cath observes, it's not at all obvious that the person is *unreasonable* (and not just very wrong) in believing that they should randomly kill people.

A perhaps more serious issue is that on many variants of RE, people may end up disagreeing even if they share the same starting point. This is because many versions of the method fail to specify exactly how one should adjust one's beliefs or credences to achieve equilibrium. Indeed, they even allow one to achieve equilibrium simply by forming the belief that one's principles and the contents of one's judgment are coherent without changing either, even if they are not (Woods 2019, 329–30). This threatens to make it completely *arbitrary* what one should believe after RE. To be sure, proponents of RE could reply that some revisions are illegitimate since they are *ad hoc*. But whether a revision is *ad hoc* itself depends on whether it is an unjustified assumption made just to save one's view, and thus something that is itself up for grabs in the pursuit of reflective equilibrium, if what determines legitimate adjustments depends on the subject's own take. A similar argument can be given if we are accused of violating some other theoretical virtue, such as elegance, in our pursuit of reflective equilibrium. RE seems simply too liberal to yield epistemic

justification, unless it is formulated so as to constrain the adjustment of initial convictions in a principled manner.

4. The Uses of Reflective Equilibrium: From Alethic Justification to Moral Understanding

We have emphasized that there are many ways to understand the method of reflective equilibrium – at least 256, if you only focus on the most important theoretical choices. The discussion of criticisms and replies in the previous section suggests that there is good reason to make certain choices. Insofar as we're interested in epistemic justification, the *subject* should be an individual, not a group (decision point A in our schema), though the fact that other people believe something can be relevant in our reasoning. The *inputs* should be considered judgments rather than unreflective intuitions (B), and because of the holistic nature of the process, whole theories or sets of principles rather than individual principles should ideally be balanced with judgments and background theories (C). Our discussion of coherentism implies that the most plausible epistemically justificatory claim to make for RE is that arriving at an equilibrium in this way *amplifies* justification that individual inputs already have (H). To the extent that RE claims the pre-existing justification is never sufficient for full belief, it still remains a distinctively, if weakly, coherentist method – which, to be sure, is a problem insofar as we think there can be moral knowledge without coherence.

The epistemic credentials of RE are somewhat boosted if it is conceived of in a less subjectivist manner than it often is (F and G). It is natural enough to think that RE requires subjects to filter out judgments that are apt to be unreliable *by their own lights*, and similarly make adjustments to the starting points that increase coherence as they see it. But this is not the only approach. RE could be *formally objectivist* in the sense that something would count as a considered judgment only if it is formed in a way that is in fact generally reliable, and coherence would be a matter of the subject's belief-contents or credences actually supporting each other. This would still

be consistent with the method being a form of RE, unlike a *substantively* objectivist view, which would require starting only from substantively correct assumptions. Notably, on such formally objectivist variants, we may be mistaken about whether our beliefs are in reflective equilibrium. If so, RE, like other epistemic methods, is not fully operationalizable in the sense that one would always be in a position to know whether one follows it (Williamson, 2008).

Reflective Equilibrium as a Zetetic Method

In spite of the refinements above, there is reasonable doubt about whether even the best form of RE suffices to guarantee that the resulting beliefs are epistemically justified, at least in any sense linked to truth-conduciveness, which we might label *alethic* justification. It is thus worth asking whether the method is better conceived of as aiming at something else (decision point I). We will set aside the option of regarding it as a purely descriptive method, since this option is of little relevance from the perspective of doing *normative* ethics (cf. Singer 1974, McPherson 2015). This still leaves two interesting possibilities: perhaps RE should be understood as a method of *inquiry that aims at moral understanding* or of *dialectical or moral justification*.

Let us start with inquiry. What do we mean by a method of inquiry? According to Jane Friedman's (2019; 2020) influential account, we can think of inquiry as consisting of three stages: forming questions, engaging in activity to answer them, and forming a belief that answers the question and concludes the inquiry. She rightly highlights that there are norms, including epistemic norms, governing all of these stages. In some circumstances, we should adopt a questioning attitude towards certain propositions, and perhaps shouldn't do so with respect to others. Importantly for our purposes, there are also norms for how to go about answering questions, which vary according to the subject matter. They come apart from norms that say which beliefs are justified, as shown by the fact that forming justified beliefs on matters that are irrelevant to and distract from an ongoing inquiry can be prohibited by the relevant norm of inquiry (Friedman 2020, 503).

We think it is very natural to conceive of RE as a *method of responsible moral inquiry*. After all, it is typically presented as a way to *figure out* which theory one should endorse, and many defenders talk about it in terms of *deliberation* (e.g. Scanlon 2002, Cath 2016). This also makes good sense of why the method is often described in first-personal terms: how should *I* or *we* go about deciding which moral principles or theory to accept?

But why would RE be the correct answer to such questions? Our suggestion is that the best case for this begins from the assumption that an important epistemic aim of moral inquiry is *understanding* rather than justified belief or knowledge.[3] It is widely agreed that to understand something, such as a machine, it does not suffice to have isolated items of knowledge about it (e.g. that such and such part prevents overheating); one must also know *why* it is the way it is and what would happen if things were otherwise (e.g. if the valve were opened, the fuel would flow again) (Kvanvig 2003, Grimm 2012). Such *objectual* understanding (understanding a subject matter) thus involves a grasp of explanatory relations and an ability to make appropriate inferences, and is a matter of degree (Kvanvig 2003). In the moral case, as Alison Hills puts it, "If you truly understand why your action is right, you are aware of the reasons why it is right, and if the situation were a bit different, you could correctly draw the conclusion that some other action would be right instead, and explain why" (Hills 2020, 409). Objectual understanding of morality in general goes beyond understanding why particular actions are right, and requires some general grasp of right- and good-making features and their relations, among many other things. It thus puts us in a position to reliably do the right thing in a wide range of possible situations and articulate for others why that is the case. Thus, it seems to have particular moral value – indeed, Aristotle may have thought it was necessary for full moral virtue (*NE* 1144b).

[3] For the link between RE and understanding in the context of philosophy of science, see Elgin 1996 and Baumberger and Brun 2021.

Why think RE is the best method of inquiry if the aim is understanding? Earlier, we suggested that you could know it is wrong to bomb innocent civilians even if your moral principles do not match this belief, so that RE is not necessary for justification. Be that as it may, in such a case you won't understand *why* it is wrong to bomb civilians – you might even find it mysterious by your own lights. Nor will you be able to tell in a principled way how things would have to differ for bombing to be morally permissible (say, the civilians would have to be responsible for ordering a genocide), because your principles don't discriminate between the present case and alternatives in this respect (since they allow bombing anyway). In contrast, if your moral outlook is in (well-founded) reflective equilibrium, or close to it, you will be able to tell why bombing in this sort of case is wrong, and when it might not be, by drawing on your principles, and possibly background theories. It is no wonder that on many conceptions of objectual understanding, it involves a grasp of "relevant coherence-making relations between propositions comprising some subject matter" (Carter and Gordon 2014, 7).

To be sure, since RE might not lead to true beliefs (not to mention knowledge), it won't *guarantee* understanding either, insofar as it is factive. But then again, nothing will. Our more modest claim is just that RE is the *best feasible method* for achieving moral understanding. Perhaps in principle you could be gifted with a grasp of all the correct moral principles, their rationales, and their consequences for every conceivable situation. But since understanding involves not only having the beliefs you ought to have but also seeing how they hang together and being able to reason about possibilities, it is very hard to see how it could realistically be achieved without engaging in the method of RE. After all, RE involves reflecting on why our considered judgments would be true in the light of principles and vice versa, making adjustments to ensure the most credible contents fit in with the rest of what we believe, considering arguments for principles, and so on. While we've emphasized that there are alternatives to RE when it comes to alethic justification – indeed, some other method might be superior if your aims are narrowly alethic –

there may well not be when it comes to achieving moral understanding, given our human limitations.

Of course, focusing on RE as a method of inquiry rather than justification wouldn't make a difference if the norms for the two would not diverge. But it is very plausible that they do: you could end up with justified beliefs spontaneously, without conforming to the norms of inquiry, and you could inquire well but nevertheless end up with unjustified beliefs. This last point assumes that just as there is a gap between justified belief and true belief, there is a gap between beliefs that result from epistemically responsible inquiry and justified beliefs. To see why this is plausible, it is good to compare two people who end up with epistemically unjustified beliefs. Consider Kelly and McGrath's case of someone, say Anna, who begins RE, inter alia, from the perverse considered judgment "One is morally required to occasionally kill randomly" (2010, 347). Let us grant, for the sake of argument, that this belief survives RE (as unlikely as that is), but is not epistemically justified. And let's compare Anna with another, Bella, who holds the same epistemically unjustified belief without going through a process of RE. Is there something to be said for Anna from an epistemic perspective?

Yes, there is. While Anna, like Bella, ends up with a false and unjustified moral belief, she has done her best to subject it to epistemic scrutiny. She has considered whether her belief has been formed under circumstances that are likely to be unreliable, and moreover will have had the right idea about which circumstances are of this sort (as the objectivist interpretation of filtering has it). She has tested it against other considered convictions and candidate principles that are supported by arguments drawn from background theories she accepts, and found that it's a part of the most coherent story. (Clearly, for her to arrive at this conclusion, we must attribute to her many other false considered judgments and ill-informed background theories – that is, very bad epistemic luck.) What more could she have done to ensure that she ends up with correct moral beliefs and understands why they are correct? The *activities* that constitute her inquiry seem impeccable, even

though she ends up with *beliefs* that are (by assumption) unjustified. Even if it is fitting to target her for the epistemic analogue of blame – roughly, as one of us has argued, reducing trust in her in virtue of her epistemic character, as it is displayed in her belief (Kauppinen 2018; Kauppinen forthcoming) – her belief-forming method is not subject to such critical responses.

What could be said against this *zetetic* (inquiry-related) *interpretation* of RE? Perhaps the best objection is arguing for an alternative method of inquiry – instead of asking what *more* Anna could have done, we can ask what *else* she could have done. And there is a clear alternative, which is the one Singer (1974) proposed: Anna could have bracketed all her moral judgments and tried to figure out which moral principles are self-evident. This is the kind of *top-down strategy* notably favored by some utilitarians, who are perfectly aware of the fact that their view has counterintuitive implications. We submit that this is *not* a responsible way to conduct inquiry into ethical matters. Starting from abstract principles that seem true on reflection can result in abhorrent moral views just as well as starting from perverse considered judgments – indeed, we would claim that this is far more likely, since someone inquiring in this manner is by definition unmoored from commonsense considerations (cf. Ross 1930, 40). The same goes, *mutatis mutandis,* for proceeding by transcendental arguments that don't give weight to considered normative judgments.

Reflective Equilibrium and Dialectical Justification

Finally, let's return to justification. We've expressed scepticism about whether coherence is necessary or sufficient for alethic justification, whether it is a matter of being epistemically blameless or meeting epistemic obligations. But while dominant, this is not the only conception of justification. After all, we also talk about justification as something we can *give* to someone, and justifying as something we can *do*. This *dialectical* conception of justification as something we can offer *to* another in support of our views has venerable roots in Plato and Aristotle. Some, like Alison Hills (2009), take being able to justify our beliefs to others in this sense as partially

constitutive of moral understanding, and even if we think it's not constitutive of it, understanding will plausibly put us in a position to justify our principles to others by appeal to their implications and grounds. To be sure, I may not be able to (successfully) justify a view to just anyone in the sense of successfully convincing them of its correctness, even if I'm justified in holding it and I present considerations that are genuinely good and sufficient reasons for holding it. For example, Galileo's belief that there are mountains on the moon was justified even though he could not justify it to Cremonini, who refused to look through the telescope. Vice versa, a cult leader might provide arguments that convince the cultists to believe that they should live according to her teachings, but that doesn't mean anyone is justified in believing that they should live that way.

While some have argued that someone is justified in believing that *p* if and only if they can defend their belief that *p* to a contextually selected person who is interested in whether *p* is the case (Lammenranta 2011, 14), we think it is more fruitful to distinguish between different senses of justification, as in the Galileo and cult leader cases. Nevertheless, the dialectical notion of justification is also practically and even morally important. Insofar as RE succeeds in yielding moral understanding, it will also put us in the best possible position to dialectically justify our views to others. And even if one ends up with beliefs that are as a matter of fact unjustified, RE does arm one for justifying them dialectically. Given that wide reflective equilibrium requires in the limit that "one seeks the conception, or plurality of conceptions, that would survive the rational consideration of all feasible conceptions and all reasonable arguments for them" (Rawls 1974-5, 8), it is not a great leap of faith to think that a subject who had achieved such a state would be able to defend her beliefs to interested parties as well as anyone can. So there is a sense in which RE does result in *dialectically justifiable* beliefs, even if it doesn't at least in its purely coherentist form result in beliefs that are likely to be *true*.

And indeed, it may well be that the early proponents of RE did have something like dialectical justification in mind. Consider what Goodman says just before introducing his version of RE:

> How do we **justify** a *de*duction? Plainly, by showing that it conforms to the general rules of deductive inference… Moreover, when a deductive argument has been **shown** to conform to the rules of logical inference, we usually **consider it justified** without going on to ask what justifies the rules. (Goodman 1955, 66; bolding ours)

While Goodman doesn't explicitly say so, the most natural reading of this passage in context is that he's talking about justifying a deduction *to* someone by *showing* them that it conforms to the general rules, in which case they will usually *consider* it justified. It is to make sense of such social practices that he then introduces the idea of mutually adjusting inference principles and accepted principles. When it comes to Rawls, we've seen his early methodological remarks are ambiguous, but in his later work he explicitly says that on his political conception, "justification is not regarded simply as valid argument from listed premises, even should these premises be true. Rather, justification is always addressed to others who disagree with us" (1985, 225). It must therefore proceed from common ground (an "overlapping consensus") among reasonable citizens whose beliefs are more or less in reflective equilibrium (Rawls 1995, 144). So insofar as RE puts us in a better position to dialectically justify our views, it may after all yield a kind of epistemic justification, even if it isn't the alethic sort of justification that coherentists and foundationalists usually focus on.

Conclusion

The thought that ethical justification always begins *in medias res*, with already existing convictions, goes back to at least Aristotle. In a famous passage in the *Nicomachean Ethics*, he says that we

should begin inquiry from what seems to be the case, in particular what he calls reputable opinions (*endoxa*) held by either everyone or most people, especially those recognized to be wise (*NE* 1145b). We should then try to resolve conflicts and puzzles that arise among the *endoxa*, including resolving ambiguities and sorting through arguments for different views, and if we succeed, "we shall have offered sufficient proof" (ibid., 120; cf. Kraut 2006). While he seems to believe that, in general, people are pretty good at forming true opinions, he emphasizes that in the case of ethics in particular, one needs to have been brought up correctly to have the right starting points for studying "what is noble and what is just" (*NE* 1095b, 6) and finding its first principles. Consequently, he does not have much hope of being able to justify dialectically the claim that, say, being just is a part of leading a successful life, to people who do not already share roughly the correct moral convictions (*NE* 1179b).

If we squint a little, we can see here both the idea that ethical inquiry is a matter of sorting out and systematizing our initial moral convictions in a process that involves constructing and assessing arguments for them, and the qualification that doing so will only lead to genuine moral understanding if enough of those starting points are close enough to truth. It would be rash to conclude that Aristotle was an early proponent of a form of the modern method of reflective equilibrium. But clearly the broad approach to moral theorizing has deep roots. In this chapter, we have aimed to clarify its nature and weigh some of its pros and cons. What we have proposed is that in spite of the many problems that RE has as an account of (alethic) epistemic justification, it may be the best feasible method we have for achieving moral understanding.[4]

References

Alston, W. P. (2006). *Beyond "Justification": Dimensions of Epistemic Evaluation*. Cornell University Press.

Aristotele (2014). *Nicomachean Ethics*. Tr. R. Crisp . Cambridge: Cambridge University Press.

[4] We are grateful to Maria Lasonen-Aarnio, Max Lewis, Giulia Luvisotto, Lilian O'Brien, Connie Rosati, and Folke Tersman for discussion and comments. Work on this chapter was funded by the Academy of Finland research project *Responsible Beliefs: Why Ethics and Epistemology Need Each Other* (325620).

Audi, R. (2003). *The Good in the Right: A Theory of Intuition and Intrinsic Value*. Princeton University Press.
Audi, R. (2013). *Moral Perception*. Princeton University Press.
Baumberger, C., & Brun, G. (2021). Reflective equilibrium and understanding. *Synthese, 198*(8), 7923-7947.
BonJour, L. (1985). *The Structure of Empirical Knowledge*. Harvard University Press.
Brandt, R. B. (1979). *A Theory of the Good and the Right*: Prometheus Books.
Brink, D. (1989). *Moral Realism and the Foundations of Ethics*. Cambridge: Cambridge University Press.
Brun, G. (2014). Reflective Equilibrium Without Intuitions? *Ethical Theory and Moral Practice 17(2)*, 237-252.
Carter, J. & Gordon, E. (2014) On Pritchard, Objectual Understanding and the Value Problem. *American Philosophical Quarterly*, *51(1)*, 1-13.
Cath, Y. (2016). Reflective Equilibrium. In H. Cappelen, T. Gendler, & J. Hawthorne (Eds.), *The Oxford Handbook of Philosophical Methodology* (pp. 213-230) Oxford: Oxford University Press.
Daniels, N. (1979). Wide reflective equilibrium and theory acceptance in ethics. *Journal of Philosophy, 76*(5), 256-282.
de Maagt, S. (2017). Reflective equilibrium and moral objectivity. *Inquiry*, 60(5), 443-465.
DePaul 1986. Reflective Equilibrium and Foundationalism *American Philosophical Quarterly* 23 (1):59 - 69.
DePaul, M. (1987). Two conceptions of coherence methods in ethics. *Mind* (384), 463-481.
DePaul, M. (2006). Intuitions in moral inquiry. In D. Copp (Ed.), *The Oxford Handbook of Ethical Theory* (pp. 595--623): Oxford University Press.
Elgin, C. Z. (1996). Considered Judgment. Princeton University Press.
Elgin, C. Z. (2014). Non-foundationalist epistemology. Holism, coherence and tenability" and "Reply to van Cleve. In M. Steup, J. Turri, & E. Sosa (Eds.), *Contemporary debates in epistemology*. Malden: Wiley.
Enoch, D. (2010). The epistemological challenge to metanormative realism: how best to understand it, and how to cope with it. *Philosophical Studies, 148*(3), 413-438.
Feldman, R. (2003). *Epistemology*. Upper Saddle River, NJ: Prentice-Hall.
Friedman, J. (2019). Inquiry and Belief. *Nous, 53*(2), 296-315.
Friedman, J. (2020). The Epistemic and the Zetetic. *Philosophical Review, 129*(4), 501-536.
Goodman, N. (1955). *Fact, Fiction, and Forecast*. MA: Harvard University Press.
Greene, J. D. et al. (2009). Pushing moral buttons: The interaction between personal force and intention in moral judgment. *Cognition 111(3)*, 364-371.
Greenspan, P. (2015). Confabulating the Truth: In Defense of "Defensive" Moral Reasoning. *The Journal of Ethics, 19*(2), 105-123.
Grimm, S. (2012). The Value of Understanding. *Philosophy Compass, 7*(2), 103-117.
Haack, S. (1993). Double-aspect foundherentism: A new theory of empirical justification. *Philosophy and Phenomenological Research, 53*(1), 113-128.
Hills, A. (2009). Moral testimony and moral epistemology. *Ethics, 120*(1), 94-127.
Hills, A. (2020). Moral Testimony: Transmission Versus Propagation. *Philosophy and Phenomenological Research, 101*(2), 399-414.
Hirvelä, J. (2022). Justification and the knowledge-connection. *Philosophical Studies 179*, 1973-1995.
Joyce, R. (2006). *The Evolution of Morality*. Cambridge, Mass: MIT Press.
Kagan, Shelly (1998). *Normative Ethics*. Boulder, CO: Westview Press.
Kant, I. (1785). *Groundwork of the Metaphysic of Morals*.

Kauppinen, A. (2007). The rise and fall of experimental philosophy. *Philosophical Explorations, 10* (2), 95 – 118.
Kauppinen, A. (2018). Epistemic Norms and Epistemic Accountability. *Philosophers' Imprint, 18*.
Kauppinen, A. (2019) Who's Afraid of Trolleys? In Jussi Suikkanen & Antti Kauppinen (eds.), *Methodology and Moral Philosophy*. Routledge.
Kauppinen, A. (forthcoming). The Epistemic vs. the Practical. *Oxford Studies in Metaethics* vol. 18.
Kelly, T., & McGrath, S. (2010). Is reflective equilibrium enough? *Philosophical Perspectives, 24*(1), 325-359.
Kitcher, P. (2011). *The Ethical Project*. Harvard University Press.
Korsgaard, C. (1996). *The Sources of Normativity*. Cambridge University Press.
Kraut, R. (2006). How to justify ethical propositions : Aristotle's method. In R. Kraut (Ed.), The Blackwell Guide to Aristotle's Nicomachean Ethics (pp. 76--95). Blackwell.
Kvanvig, J. L. (2003). *The Value of Knowledge and the Pursuit of Understanding*. Cambridge University Press.
Lammenranta, M. (2011). Disagreement, Skepticism, and the Dialectical Conception of Justification. *International Journal for the Study of Skepticism, 1*, 3-17.
Lemos, N. (2019). Lemos, Foundationalism and Coherentism in Moral Epistemology. In A. Zimmerman, K. Jones, & M. Timmons (Eds.), The Routledge Handbook of Moral Epistemology. New York: Routledge.
Lewis, C. I. (1946). *An Analysis of Knowledge and Valuation*. LaSalle: Open Court.
Littlejohn, C. (2012). *Justification and the Truth Connection*. Cambridge University Press.
Littlejohn, C. (forthcoming). A Plea for Epistemic Excuses. In F. D. Julien Dutant (Ed.), *The New Evil Demon Problem*: Oxford University Press.
McPherson, T. (2015). The Methodological Irrelevance of Reflective Equilibrium. In C. Daly (Ed.), *The Palgrave Handbook of Philosophical Methods* (pp. 652-674): Palgrave Macmillan.
Mikhail, J. (2011). *Elements of Moral Cognition: Rawls' Linguistic Analogy and the Cognitive Science of Moral and Legal Judgment*: Cambridge University Press.
Nozick, R. (1974). *Anarchy, State, and Utopia*. New York: Basic Books.
Olsson, E. J. (2002). What is the problem of coherence and truth? *Journal of Philosophy, 99*(5), 246-272.
Parfit, D. (1984). *Reasons and Persons*. Oxford: Oxford University Press.
Paulo, N. (2020) The Unreliable Intuitions Objection Against Reflective Equilibrium. *The Journal of Ethics 24 (3)*, 333-353.
Petrinovich, L. F. and O'Neill, P. (1996). Influence of wording and framing effects on moral intuitions. *Ethology and Sociobiology* 17 (3), 145-171.
Pollock, J. (1986). *Contemporary Theories of Knowledge*. Rowman & Littlefield.
Pust, J. (2000). *Intuitions as Evidence*. Routledge.
Rawls, J. (1951). Outline of a Decision Procedure for Ethics. *Philosophical Review 60*(2), 177-197).
Rawls, J. (1971). *A Theory of Justice: Original Edition*. Belknap Press.
Rawls, J. (1974). The Independence of Moral Theory. *Proceedings and Addresses of the American Philosophical Association, 48*, 5 - 22.
Rawls, J. (1985). Justice as fairness: Political not metaphysical. *Philosophy and Public Affairs*, 14(3), 223-251.
Rawls, J. (1993). *Political Liberalism*. New York: Columbia University Press.
Rawls, J. (1995). Reply to Habermas. *Journal of Philosophy*, 92(3), 132-180.
Rawls, J. (2001). *Justice as Fairness: A Restatement*. Harvard University Press.
Ross, W. D. (1930). The Right and the Good. *Philosophy, 6*(22), 236-240.
Sayre-McCord, G. (1996). Coherentist Epistemology and Moral Theory. In W. Sinnott-Armstrong & M. Timmons (Eds.), *Moral Knowledge? New Readings in Moral Epistemology*. Oxford: Oxford University Press.

Scanlon, T. M. (2002). Rawls on Justification. In S. R. Freeman (Ed.), *The Cambridge Companion to Rawls* (pp. 139). Cambridge University Press.
Schnall S, Haidt J, Clore GL, Jordan, A. H. (2008) Disgust as Embodied Moral Judgment. *Personality and Social Psychology Bulletin*, *34(8)*,1096-1109.
Schwitzgebel, E. & Cushman, F. (2012) Expertise in Moral Reasoning? Order Effects on Moral Judgment in Professional Philosophers and Non-Philosophers. *Mind and Language 27* (2),135-153.
Sidgwick, H. (1907). *The Methods of Ethics*. London: Macmillian.
Singer, P. (1974). Sidgwick and Reflective Equilibrium. *The Monist, 58*(3), 490-517.

Srinivasan, A. (2020). Radical Externalism. *Philosophical Review, 129*(3), 395-431.
Street, S. (2006). A Darwinian Dilemma For Realist Theories of Value. *Philosophical Studies, 127*, 109-166.
Tersman, F. (1993). *Reflective Equilibrium an Essay in Moral Epistemology*. Coronet Books.
Tersman, Folke (2008). The reliability of moral intuitions: A challenge from neuroscience. *Australasian Journal of Philosophy* 86 (3):389 – 405.
Tersman, F. (2018). Recent work on reflective equilibrium and method in ethics. *Philosophy Compass, 13*(6), 1-10.
Tomasello, M. (2016). *A Natural History of Human Morality*: Harvard University Press.
van Cleve, J. (2011). Can Coherence Generate Warrant Ex Nihilo? Probability and the Logic of Concurring Witnesses. *82*(2), 337-380.
van Parijs (1991), Why surfers should be fed: The liberal case for an unconditional basic income. *Philosophy and Public Affairs* 20 (2),101-131.
van Roojen, M. (2015). Moral intuitionism, experiments and skeptical arguments. In Anthony Booth & Darrell Rowbottom (eds.), *Intuitions*. Oxford University Press.
Vavova, K. (2021). The Limits of Rational Belief Revision: A Dilemma for the Darwinian Debunker. *Nous, 55*(3), 717-734.
Wedgwood, R. (2006). How we know what ought to be. *Proceedings of the Aristotelian Society, 106*(1), 61–84.
Williamson, T. (2008). Why epistemology cannot be operationalized. In Q. Smith (Ed.), *Epistemology: New Essays*. Oxford: Oxford University Press.
Williamson, T. (forthcoming). Justifications, Excuses, and Sceptical Scenarios. In F. Dorsch & J. Dutant (Eds.), *The New Evil Demon*. Oxford: Oxford University Press.
Woods, J. (2019). Against Reflective Equilibrium for Logical Theorizing. *Australasian Journal of Logic, 16*(7), 319.