Synthese (2024) 203:79
https://doi.org/10.1007/s11229-024-04495-1

ORIGINAL RESEARCH

Is there a defensible conception of reflective equilibrium?

Claus Beisbart[1] · **Georg Brun**[1]

Received: 18 February 2023 / Accepted: 5 January 2024 / Published online: 22 February 2024

Abstract
The goal of this paper is to re-assess reflective equilibrium ("RE"). We ask whether there is a conception of RE that can be defended against the various objections that have been raised against RE in the literature. To answer this question, we provide a systematic overview of the main objections, and for each objection, we investigate why it looks plausible, on what standard or expectation it is based, how it can be answered and which features RE must have to meet the objection. We find that there is a conception of RE that promises to withstand all objections. However, this conception has some features that may be unexpected: it aims at a justification that is tailored to understanding and it is neither tied to intuitions nor does it imply coherentism. We conclude by pointing out a cluster of questions we think RE theorists should pay more attention to.

Keywords Coherence · Epistemic justification · Reflective equilibrium

1 Introduction

Reflective equilibrium ("RE" for short) is by now well-known in philosophy. It was first proposed by Goodman (1983) as an answer to the question of how to justify (deductive and inductive) inferences and principles of inference. Later, Rawls (1999), to whom we owe the term "reflective equilibrium", introduced the RE into ethics to justify his theory of justice. Since then, RE has been popular in ethics (e.g. Daniels, 1996; DePaul, 1998; 2011; Scanlon, 2014), its use has been extensively discussed in rationality theory (Cohen, 1981; Stein, 1996), and some authors have

✉ Georg Brun
Georg.Brun@unibe.ch

Claus Beisbart
Claus.Beisbart@unibe.ch

[1] Institute for Philosophy, University of Bern, Länggassstrasse 49a, Bern 3012, Switzerland

suggested RE more generally for philosophical theorizing (Lewis, 1983:x; Keefe, 2000:ch 2.1) or as an account of epistemic justification (Elgin, 1983; 1996; 2017). But over the years, RE was also challenged by a considerable number of objections.

One may therefore think that the power, but also the problems and limitations, of RE have been adequately discussed. This, however, is not the case. RE continues to be controversial. A main reason for this is that defenders and detractors have too often relied on rather sketchy ideas of RE – if they have targeted RE at all, rather than positions to which they take RE to be wedded to, for example, coherentism or intuitionism. Another problem is that a number of different versions or specifications of RE have been proposed (by the authors referenced above and many others). Since some objections are fatal for one version but ineffective against another, it is difficult to assess how defensible RE is overall. Additionally, there is a danger that proponents of RE switch between different versions of RE when they try to defend it against several objections.

This paper is meant to be a reaction to this rather problematic state of the discussion. The overarching goal is to address the question whether there is a consistent conception of RE that can be defended against the most prominent objections. To answer this question, we first provide (in Sect. 2) a systematic overview and analysis of the critical arguments. For each objection, we investigate why it looks plausible, on what standard or expectation it is based, and how convincing it is. We further assess responses to the objections and explore the consequences for a defensible conception of RE.[1] In Sect. 3, we draw the results together and argue that there is a version of RE that promises to withstand all objections. The outcome is a version of RE that may look rather surprising, because it differs in several respects from ideas often associated with RE. We conclude in Sect. 4 by pointing to open questions that RE theorists should answer.

As a basis for our investigation, we need a pre-conception of RE. We assume that RE aims at epistemic justification. Although RE has been described in a variety of ways, there are two key ideas which constitute the core of RE and refer to static and dynamic aspects, respectively. First, justification is a matter of agreement between commitments (often called "judgements") and a systematic set of theoretical principles; second, such an agreement can be reached by a process in which the commitments and the principles are mutually adjusted to each other. To elaborate a bit, as a *process of equilibration*, RE starts with an agent's initial commitments about a certain subject matter. The agent then introduces principles which are intended to form a theory of the subject matter and to account for her commitments. If there are (as it is most likely) discrepancies between the proposed principles and the agent's commitments, the agent has to iteratively adjust principles and commitments until she ideally reaches a state in which her commitments agree with the theory and are supported by background theories (as required by so-called "wide" RE). In the resulting *state of reflective equilibrium* ("RE state" for short), both the commitments and the theory

[1] Our analysis differs from the available surveys of objections against RE (e.g. Cath, 2016; Daniels, 2018; Knight, 2023) by being more comprehensive and systematic, and also because we do not only aim at answering the objections, but also at analysing the expectations that drive the objections and at identifying desiderata for an account of RE that can answer all important objections.

(the "position", for short) are justified. To simplify matters, we assume that commitments have propositional content and that RE is applied by one individual epistemic agent. We thus bracket issues that arise within groups, for example, how a group identifies its initial commitments and how it proceeds in the process of equilibration. We also often gloss over the fact that being in RE is a matter of degrees (Baumberger & Brun, 2017).[2]

Before we start, we should clarify our project with four remarks. First, in this paper, we ask whether there is a conception of RE that can be defended against all objections, but not whether there is only one such conception. There may be other versions of RE that meet the objections, but analysing alternative conceptions of RE is not in the scope of this paper. Accordingly, there are descriptions of RE that we will not discuss.

Second, the objections against RE presuppose certain expectations, although this is seldom made explicit. For instance, is it a problem if reaching RE is not sufficient or not necessary for obtaining justification? The answer depends on whether one expects RE to be sufficient or necessary for justification. In our analysis of the objections, we will thus try to make clear what expectations they rely on. Note that one can defend RE against an objection by rejecting an expectation behind it. But if this strategy is applied too often, we obtain a 'defence' of RE that is hardly of interest. We thus aim at a conception of RE which meets interesting and plausible expectations, although we are critical of some expectations that have been connected to RE. Consequently, this paper is also about the expectations that one may reasonably invest in RE.

Third, some readers may already want to disagree with our pre-conception of RE. But since our goal is only to find one version of RE that can be defended against the objections, we need not show that our pre-conception is the only reasonable one. Furthermore, if we can obtain such a defensible version from our pre-conception, this is some support for this pre-conception – others may not lead us that far.

Fourth, there is a huge literature with objections against RE and we cannot address each and every objection separately in this paper. Nor is this necessary for our purpose since most objections can be seen as variations on a few recurrent themes. Our proposal for systematizing the objections can be gathered from the structure of Sect. 2: each subsection addresses what we take to be a major type of objection. We present one or a few representative sources, often giving priority to early or particularly clear versions. No attempt has been made to find the earliest, let alone all, sources for any type of objection.

[2] Two remarks on terminology. First, "reflective equilibrium" may refer to a target state and/or to the process to get there. For clarity, we often call the target state "RE state." Second, the literature commonly speaks of RE as a "method". We avoid this characterization since we do not want to imply that descriptions of RE processes provide recipes for actually carrying out an investigation (see Rechnitzer, 2022 for a case study of inquiries explicitly conducted as RE processes). Instead, the process may be understood as a reconstruction of an actual course of inquiry, specifying how a RE state could have been reached by a sequence of adjustments (e.g. Elgin, 2017:64; Tersman, 1993:15).

2 Analysis and discussion of major objections against RE

2.1 General objection: RE is uninformative

A very general objection is that RE is an uninformative account of justification, insofar as every reasonable course of inquiry can be considered to be an application of RE. Basically, RE instructs the epistemic agent to consider all the reasons she has for and against various claims and to resolve conflicts in the way she judges best – but what else may she do? RE remains silent on the epistemologically crucial questions of what should count as a reason in what kind of inquiry, what exactly a conflict is and how it should be resolved. We are not told, for example, under which conditions intuitions are trustworthy – so the objection goes (Foley, 1993:128; Williamson, 2007:244–6).

One reason why one might find this charge plausible is that some prominent descriptions of RE are extremely unspecific. In Rawls's writings after *A Theory of Justice*, for example, one finds a tendency to dilute the key idea of RE when he describes the target state as one that "would survive the rational consideration of all feasible conceptions and all reasonable arguments for them" (1975:8; see also 2001:31). This threatens to turn RE into the uninformative thought that we should adopt whatever seems the best view all things considered. Such an impression may also be fuelled by the claim, defended by DePaul (1998) and Scanlon (2003), that there is no reasonable alternative to RE.

The expectation behind this objection is that RE should be a distinctive account of epistemic justification. It should neither boil down to epistemic *anything goes* nor to mere platitudes about inquiry. Rather, it should help determine which inquiries count as reasonable. Consequently, an appropriate way of answering the objection is to develop a more specific conception of RE by addressing questions such as: What strategies of adjustments should be chosen if commitments conflict with theories? What does it mean to propose *systematic* principles? As a matter of fact, the literature offers conceptions of RE which answer such questions and thus give informative accounts of justification. Most noticeably, virtually all versions of RE, how sketchy ever they are, exclude many forms of foundationalism. Since in principle every commitment and every element of a theory is revisable, even wide RE excludes "indubitable starting points" (contra Singer, 2005:347); and since RE provides the standard for justification, no commitment can be justified if it is not part of a RE state.[3] The fact that proponents of RE agree on this clearly shows that RE is not completely uninformative. And many defenders of RE go further in specifying RE, which then becomes incompatible with additional accounts of justification, for example, pure coherentism (see Sect. 2.5). Since this paper goes in a similar direction and aims at elaborating the basic aspects of a more specific conception of RE, our project as a whole can be seen as an answer to the objection that RE is uninformative.

Although we agree with the expectation that RE should be a distinctive account of justification, we also want to insist that we must not expect RE to answer all kinds

[3] In terms of BonJour's (1985:26) well-known classification, RE thus excludes "strong" and "moderate" foundationalism (see Sect. 2.5 for further discussion).

of epistemological questions. We take it that the aim of RE theory is to clarify the general structure of epistemic justification. A theory of RE should thus specify what the components of a justified position are (e.g. commitments and theories), what features the components should have and how they should relate to each other. But it need not, indeed should not, fulfil Singer's request to determine which moral intuitions (if any) make for credible commitments (see also Tersman, 2008). This request cannot be answered by a general theory of epistemic justification since one cannot determine *a priori* particular factors of justification and decide *a priori* on the exact justificatory role of, for example, perception (Walden, 2013; see also Scanlon, 2003). In this respect, RE is like other general theories of epistemic justification such as reliabilism. As is well-known, reliabilism takes justification to be a matter of belief formation through reliable processes without spelling out which particular processes are reliable.

2.2 Input: intuitions are problematic

A second objection takes RE to be unacceptable as an account of justification, because (1) RE relies on intuitions as input, and (2) intuitions are problematic.[4] This is probably the most prominent objection to RE, and it comes in many versions. Very early on, Rawls was challenged by philosophers who claimed that our moral intuitions are typically shaped by prejudice, religious education, unreflected dogma and other factors that can lead us to intuitive convictions which are deeply problematic from a moral point of view (Hare, 1973; Singer, 1974). More recently, similar worries have been backed by appeals to empirical findings from neuroscience and evolutionary psychology (Singer, 2005) or from experimental philosophy (Appiah, 2008). Targeting the very idea of a RE, not only its application in ethics, some experimental philosophers have claimed that RE implies "Intuition-Driven Romanticism" (Weinberg et al., 2008:20), which faces severe challenges from, for example, cultural variations in intuitions.

Undeniably, these worries must be taken seriously. As premise (1) claims, RE is often conceived as essentially involving intuitions, for example, when the commitments entering a RE process are identified with intuitions or when RE is straightforwardly described in terms of intuitions (e.g. DePaul, 2006). As far as (2) is concerned, intuitions can obviously be incorrect, unjustified or misleading.

The crucial expectation behind this objection is that a process of justification should not use problematic input (more on how this expectation may be motivated in the next sections). But even if we accept this expectation, the objection be can countered. We proceed in two steps. We first discuss premise (1) and argue that RE should not be characterized in terms of intuitions. We secondly turn to a popular attempt to reject the whole argument, which will lead us to our own reaction to the objection.

A strong reading of (1) identifies the input to RE with the agent's intuitions on the topic under consideration. In this strong sense, (1) should be rejected. On the one hand, input limited to intuitions is too narrow since it unduly excludes commitments from sources such as memory or perception and, most importantly, commitments that

[4] See Brun, 2014 for an extended discussion of many points mentioned in this subsection.

have been explicitly inferred from a theory that the agent has developed in an earlier RE process. Such commitments are not intuitions since it is agreed that intuitions are not explicitly inferred, but there is no reason to generally exclude such commitments as input. On the other hand, input that includes all the agent's intuitions is too wide since the agent may have a 'recalcitrant' intuition that p to which she is not committed since she actually considers p to be false. Clearly, such an intuition should not enter the RE process.

Proponents of RE are thus well-advised not to base their characterization of RE on the notion of intuition. They do not even need to assume that there are any intuitions in the sense in which they are understood by recent theories of intuitions (see Burkard, 2012 for an overview).

Still, what we have said so far is compatible with weaker versions of (1), which hold that intuitions are needed in specific fields of inquiry. For instance, it may turn out that appealing to intuitions is needed in ethics or in logic. The conjecture that intuitions are necessary in some domains is supported by the observation that even opponents of RE seem to resort to intuitions. Singer (2005:351), for example, insists that some basic moral principles are "rational intuitions"; Stich and Nisbett (1980:198) clearly appeal to intuitions when they claim that it is "completely obvious" that adherents of the gambler's fallacy rely on a "patently" invalid rule.

The result of our discussion of (1) thus is that a general, not domain-specific, characterization of RE should avoid reference to intuitions, but that intuitions may play an important role as input to RE processes in inquiries about some specific topics. The problem then is that these intuitions can be problematic, as emphasized in (2). This brings us to the second step of our counter.

The most popular response to the objection from problematic intuitions is the requirement that intuitions be 'filtered' before they enter RE (usually this is just a special case of the general requirement that the input to RE be *considered* judgements). For instance, an intuition may be allowed to enter the RE process only if it was formed without hesitation and with strong confidence (Rawls, 1999:42) or if it is based on adequate information (Daniels, 1979:258). Independent of how the filter works in detail, the response avoids the objection by saying that RE relies (at most) on filtered intuitions (modification of premise 1) and that filtered intuitions are not problematic (denial of premise 2 for filtered intuitions).

The filter response, however, comes with its own problems. If RE is applied to moral theories, one might object that some of Rawls's filters (e.g. excluding commitments made "when we are upset"; 1999:42) beg moral questions (Daniels, 1979:258n3). The general problem is how the proposed filters can be given an epistemic justification.

This way of framing the problem with filters invites a different thought, which leads to our reply to the objection. Sorting out problematic intuitions is a matter of substantial epistemic considerations, which often also involve empirical questions. Thus, it is not the business of a theory of RE to decide under which conditions (if ever) intuitions are unproblematic. This decision should be left to theories of intuitions, which will then be incorporated in a RE process as background theories. We propose to proceed in the same manner with other sources of commitments such as perception, memory and testimony. Developing epistemic theories of such sources

is not a matter of a general theory of epistemic justification, but calls for a critical epistemology of perception, memory etc.[5] The crucial move then is to take premise (2) seriously without taking a general stance on the reliability of intuitions (or of perception, memory etc.): specific epistemologies about these sources of commitments should be incorporated into RE processes as background theories and these theories should be instrumental in deciding which commitments are rejected as problematic in the RE process. This does not imply that such background theories need to be tuned to RE specifically. It is sufficient if they tell under which conditions and to what extent commitments based upon intuition, memory etc. have some credibility that obtains independent of the RE process under consideration, but is relevant to the overarching epistemic objective at which justification is aimed (we come back to this aim in Sect. 2.6, and to independent credibility in Sect. 2.5).

All in all, defenders of wide RE can deny the general claim that intuitions are an essential part of RE; they can insist that those RE processes which use intuitions as input commitments need to include a background theory that permits a critical evaluation of the intuitions; and they can impose an analogous requirement on other sources of commitments. This is what we will do for the remainder of this paper, adopting a strategy suggested by DePaul (2011:lxxxix) as a further development of Daniels's original conception of wide RE. In a similar vein, Tersman (2008; 2018) appeals to second-order beliefs about the reliability of intuitive commitments. The most explicit and radical step in this direction can be found in Elgin's (2017:ch. 4) theory of RE. For her, the primary object of epistemic justification is what she calls an *account*, which includes not only commitments to propositions about a subject matter but also about the standards needed to evaluate such commitments.

Of course, background theories or second-order beliefs about the reliability of certain sources need their own justification, and for proponents of RE, this justification is also a matter of a RE, one that aims at justifying an epistemic theory about a certain source of commitments. It is attractive to think that this justification is combined with the justification of the relevant commitments themselves. This would mean that justification by RE is ultimately holistic (Elgin, 2017:ch. 4; see also DePaul, 2013:4473) even if one inquiry can only proceed by addressing one specific subject matter.

Our reply to the objection from problematic intuitions did not take issue with the expectation behind it, namely that a process of justification must not use problematic input. But it is not clear whether we should grant this expectation. Since the application of RE generates some pressure to revise the commitments that entered it as input, there is also the option to argue that the influence of problematic intuitions is obliterated through the RE process. Since this idea is important in answering the next two objections, we discuss it in this context.

[5] Singer attacks the idea that wide RE permits to reject any moral intuition deemed problematic by a background theory. Without giving moral intuitions a special status – that is, a certain degree of immunity from revision – wide RE is not a distinctive account of justification any more, Singer argues (2005:347). This worry has been addressed in the preceding subsection.

2.3 Process: garbage in, garbage out

The second objection focused on intuitions and thus a special kind of input. But this focus may be a distraction. Whatever the sources of the input commitments may be, there is no denying that people hold problematic commitments.[6] Not only will moral commitments sometimes be shaped by problematic traditions ("animals have no moral standing") or reflect prejudice ("homosexuality is unnatural"); also commitments based upon perception can be distorted (the moon illusion), strong mathematical intuitions may be false (the gambler's fallacy), logical (and many other) commitments are subject to biases (Wason selection task), and so on. Since RE crucially depends on input commitments – so the garbage-in-garbage-out objection goes – RE will allow us to 'justify' claims known to be false. Stich and Nisbett (1980), for example, have argued that some people may hold commitments that lead them into adopting a principle that licences the gambler's fallacy. For the proponents of RE, Stich and Nisbett claim, these people would have to be justified in their commitment to the gambler's fallacy as long as their principle is in equilibrium with their commitments.

The crucial expectation behind this objection is, very roughly, that RE should yield acceptable results, quite independent of the quality of the input. In some readings, this expectation is clearly too strong (Kelly & McGrath, 2010). For instance, objectors must not expect that a RE state contains only truths. RE is an account of justification, and justification does not guarantee truth. It is also inappropriate to expect that the RE never leads to theories or commitments that we find unjustified. Justification is defeasible and relative to an epistemic agent. A proposition p can be justified for an agent given her epistemic position even if it is not justified to other people and even if other people know that p is false. Likewise, a proposition can be justified to an agent at some time even if it is not justified to her at an earlier or later time. Thus, Stich seems to demand too much when he objects that even wide RE fails to make it "impossible" to arrive at a "set of principles and convictions that includes some quite daffy inferential rule" (1990:85–6). More generally, it is inappropriate to expect that a process of justification should perform epistemic alchemy by turning arbitrary input into sensible results ("garbage in, gold out"). In sum, the objection can rebutted if it is based on such inappropriate expectations.

If the expectations are toned down, a more plausible version of the objection results since it is reasonable to expect that applying an account of justification leads the epistemic agent to detect and to correct or give up commitments that are unjustified to her given her epistemic position. A full answer to the objection thus needs to show that RE has the power to identify and revise problematic input. In this respect, RE does indeed contain a number of resources. First, RE requires coherence and therefore the epistemic agent is under pressure to revise commitments or principles, not only if they contradict or undermine each other, but also if they do not sufficiently support each other. Second, *wide* RE requires that the commitments and the theory be sup-

[6] Paulo, 2020 makes this move explicitly in reaction to the arguments we presented in Sect. 2.2.

ported by background theories that are in turn justified (e.g. using a different RE).[7] In the case of the gambler's fallacy, for example, probability theory is a relevant background theory which provides the resources for rejecting the fallacy for people who are justified in accepting probability theory (Elgin, 1996:118–9). Furthermore, as mentioned in Sect. 2.2, epistemological background theories can provide the means to identify and remove commitments that have been formed in problematic ways.

Given these resources, the RE process will put pressure on the agent to revise problematic commitments and subsequently develop a RE state that does not contain a theory that is clearly unjustified. In fact, many people who were initially attracted by the gambler's fallacy have corrected themselves due to such pressure. However, the question remains whether RE's critical potential for dealing with problematic input is strong enough. We address this issue in the next Section on the objection from conservativity.

Before we do that, a final point about the garbage-in-garbage-out objection must be clarified. RE processes are not guaranteed to end in RE states. It may well be that a certain input does not permit the agent to finish the process at all: garbage in, nothing out. In such a situation, no (strong) justification has been established and problematic commitments still present at this stage do not count against RE.

2.4 Process: RE is too conservative

Conservatism objections to RE come in various versions. The strongest versions hold that RE limits revisions of commitments in some principled way. This complaint is often related to Rawls's statement that "there is a definite if limited class of facts against which conjectured principles can be checked, namely, our considered judgments in reflective equilibrium" (Rawls, 1999:44). A first version of the objection interprets Rawls as claiming that there is a fixed set of judgements which can be used to test proposed theories and that these judgements are therefore immune from revision in light of the theory (Hare, 1973:145–6). A second version of the objection has been prompted by Rawls's characterization of RE as aiming to describe a person's moral sense in analogy to a Chomskyan competence (Rawls, 1999:41–2). Read in this way, RE will permit us to revise a commitment only if it can be classified as a performance error (Singer, 1974). We put this second version aside since we do not understand RE as a method for developing an account of some competence and therefore see no reason to limit revisions to corrections for performance errors.[8]

The conservativity objection (in the first version as well as in the versions discussed below) makes sense only if we expect that applying an account of justification must potentially lead to more or less thorough-going revisions of one's position. We do not take issue with this expectation since it certainly makes an account of justifica-

[7] Sometimes (see, e.g. Haslanger, 1999:465–6 on Stich), the garbage-in-garbage-out objection is clearly targeted at narrow RE or even more specifically at using RE for developing an account of a competence (in Chomsky's sense). For references, see the next subsection.

[8] The interpretation of RE as a method for developing an account of some competence has been extensively discussed in a debate sparked by Cohen (1981; see *The Behavioral and Brain Sciences* 4 (1981), 317–370, and 6 (1983), 487–533). It has been defended as an interpretation of Rawls by Mikhail (2010; 2011).

tion attractive if it has the potential to initiate extensive revisions. After all, examples of deeply ingrained problematic commitments in need of revision are all too easy to come by – ranging from sexist convictions to wrongheaded assumptions about the climate.

The first version of the conservativity objection can be countered by the standard argument that the corrective potential of RE must not be underestimated (e.g. Daniels, 1979:267; Tersman, 1993:49–50). As long as an epistemic agent's input commitments are not coherent and well-supported by background theories, RE puts considerable pressure on the agent to revise some commitments. In particular, no commitment (be it an input commitment or a commitment that was derived from the theory) is exempted from revision in light of the systematic principles or some background theory.

Other versions of the conservativity objection start with the observation that the process of equilibration is strongly driven by the goal of coherence. The worry then is that RE requires us to stick more or less tightly to the input commitments, because it demands revisions only if they are necessary to establish consistency or coherence among those commitments (e.g. Huemer, 2005:117). But if revisions are only needed if they resolve conflicts and foster support relations within the set of input commitments, RE "seems to amount to no more than a re-shuffling of one's initial prejudices" (Brandt, 1985:7). The process of equilibrating seems to have no serious critical potential.

This challenge is particularly relevant to versions of RE which require that the resulting commitments must be "tethered" to initial commitments (Elgin, 1996:107, 128; 2017:64, 66) or must respect input commitments (Baumberger & Brun, 2021; Lewis, 1986:88) Here, we focus on the latter requirement, which is explicitly intended as imposing some constraints on revising input commitments (Elgin's point is discussed in Sect. 2.5).

Imposing some limits on revisions of input commitments is necessary because otherwise a state of RE could be reached in a trivial way by selecting just any theory and adapting all commitments to it. Proceeding in such a way is absurd because, without any restrictions on revising commitments, the process of equilibration might simply change the subject. To block this, we need not require that the process of adjustments minimize revisions of input commitments; it suffices to require that the revisions do not result in a change of the subject matter (see Brun, 2020 for this move). A position that denies, for example, the commitment that promises have some normative force could hardly count as a theory of promises.

Since the requirement to respect input commitments does not demand to minimize revisions, it does not have the alleged consequence of leading to conservativeness. After all, input commitments can be revised for reasons that go beyond coherence in the sense of consistency and mutual support of commitments and principles. For one thing, wide RE requires support from background theories. This is admittedly a form of coherence (call it "external"), but one that is different from what the objectors have in mind, namely the "internal" coherence of the position. For another thing, some RE theorists insist that the equilibration process is also driven by epistemic goals (Elgin, 1996, 2017). Unfortunately, many champions of RE do not sufficiently appreciate this point. While they speak of "systematic principles", they rarely discuss

what makes for the aspired systematicity.[9] Systematicity, we submit, is a matter not only of coherence but of an entire range of epistemic goals or theoretical virtues.[10] Scientific theories can instantiate a range of *general* virtues such as Kuhn's (1977) accuracy, consistency, scope of application, simplicity and fruitfulness known from the debate about theory choice (see, e.g. Keas, 2018; Schindler, 2018). There are also more *domain-specific* virtues, for example, decidability in logic, identification of causal mechanisms in biology, visualizability in physics (De Regt, 2017:ch. 2.3), as well as determinacy and applicability for moral theories (Timmons, 2013; Rechnitzer, 2022:ch. 5.5). There are even project-specific epistemic goals, for example, the goal of finding a logical theory that can be directly implemented as an efficient decision procedure for classical propositional validity.

The demands that the principles constitute a systematic theory in the sense of the general and domain-specific virtues and that the principles promote project-specific epistemic goals are in fact crucial for RE. Without these demands, RE would face another triviality objection: an epistemic agent could reach a state of RE by simply declaring any coherent set of her commitments as the principles that constitute her theory. But this would not be epistemic progress; such progress is only made by striving for principles that instantiate certain virtues and fulfil certain goals. Inasmuch as the epistemic agent did not start with highly systematic and otherwise epistemically effective commitments anyway, her epistemic goals give her reason to revise commitments if she is to attain RE.[11]

2.5 Target state: coherence theories of epistemic justification are passé

Coherence does not only figure prominently in the conservativeness objection. There is also a more direct objection from coherentism. In a nutshell, the charge is that RE implies coherentism about epistemic justification and that this position faces serious problems. In this way, stock objections against coherentism are turned against RE (Fumerton, 2009; Kelly & McGrath, 2010; Lycan, 2011).

Associating RE with coherentism is plausible since all prominent defenders of RE have stressed coherence as a hallmark of RE states. Goodman, for example, speaks of a "virtuous circularity" that constitutes justification (1983:64), and for Rawls "justification is a matter of the mutual support of many considerations, of everything fitting together into one coherent view" (1999:19).

Nevertheless, the role of coherence in RE does not imply that RE amounts to a coherence theory of justification which holds that consistency and mutual support of commitments and theory suffice for justification (contra, e.g. Lyons, 1975; de Maagt, 2017; de Sousa 2010). RE requires more than coherence in this sense.[12] As explained

[9] Some remarks can be found, e.g. in Keefe, 2000:ch. 2.1 and Tersman, 1993:51.

[10] For many authors, coherence includes some of these virtues, e.g. aspects of precision (Tersman, 1993:51) or explanatory power (Thagard, 2000:ch. 3.1).

[11] It may be asked whether the goals can also be subject to revision in a RE process. For our purposes, we can leave this question open because it does not help us to address the objection under consideration.

[12] We use "coherence" in this narrow sense since this is the usual target of coherentism objections to RE. There are, of course, wider senses of coherence (e.g. Thagard, 2000) which include many aspects that we treat as additional elements of RE besides coherence in the narrow sense.

in the previous section, one additional factor is systematicity or, more generally, doing justice to epistemic goals.

Adding considerations of systematicity to coherence does not, however, suffice to secure RE from all coherentism-related attacks. In particular, one might argue that a rather implausible form of justification *ex nihilo* results if commitments can be justified using only considerations of coherence and systematicity.

The literature on RE contains a powerful but underappreciated counter against the *ex-nihilo* charge. Already Goodman (1952; *pace* Kelly & McGrath, 2010) and Scheffler (1954) have acknowledged that coherence can only boost credibility that is given independently, but cannot by itself generate credibility out of nothing. They have therefore argued that at least some commitments must have some credibility independent of their coherence with other commitments.[13] Elgin has recently elaborated this line of argument (2014:257, 267–8; 2017:ch. 4). She explicitly argues that RE cannot be reduced to coherence and admits that the role of independent credibility makes RE a form of weak foundationalism (in the sense of BonJour, 1985:26–9). The key difference to standard (Bonjour's "moderate") foundationalism is that independent credibility is never sufficient for the justification of a commitment (more exactly: sufficient for a degree of justification as it is required for knowledge or some alternative epistemic target state); for this, more is needed – according to defenders of RE: a state of RE or at least a close approximation to a state of RE.[14]

However, critics have often presented an argument that seems to show that RE cannot be given a weakly foundationalist interpretation. Weak foundationalism requires independently credible commitments, but RE relies as its input just on the commitments the agent happens to have. Hence, the weakly foundationalist interpretation of RE seems to mistakenly equate credibility with credence (Brandt, 1979:18–21; Siegel, 1992). In other words, the charge is that an agent's commitment to p signals at best that the agent *takes* p to be independently credible, but weak foundationalism requires p to actually *be* independently credible.[15]

To reply to this argument, we first note that the weakly foundationalist interpretation of RE only requires that some commitments have independent credibility *in the resulting state* since only in this state RE is supposed to constitute justification in a weakly foundationalist sense. Hence, independent credibility is only needed for some of the commitments that have survived scrutiny during the process of equilibration. As we pointed out in Sect. 2.2, epistemological background theories may imply that certain types of input commitments have or do not have independent credibility.

[13] In the literature, such independent credibility is usually called "initial" credibility since it is initial relative to the 'amplification' by coherence. In the context of RE, however, "initial" is misleading because it may refer to the start of the process. The input commitments are not exhausted by initial commitments, e.g. because, during the process, new commitments can emerge that have independent credibility without being initial input to the process.

[14] Many critical discussions of RE are not sufficiently clear about the distinction between weak foundationalism (there is independent credibility, but it does not suffice for justification) and moderate foundationalism (independent credibility can (and often does) suffice for justification); e.g. Kelly & McGrath, 2010; de Maagt, 2017. See also Cath, 2016. Note that in ethics foundationalism is often called "intuitionism" (see Brun, 2014 for an extended discussion).

[15] Kelly & McGrath, 2010 give detailed arguments why appealing to Rawls's 'filters' for considered judgements does not suffice to alleviate this worry.

Accordingly, commitments that lack independent credibility are dropped during the equilibration process.[16]

Secondly, insofar as the argument against the weak foundationalist interpretation of RE specifically targets input commitments, it can be read as the following challenge: why should we think that the epistemic agent's commitments are the right input, given that the RE process is supposed to reach a target state with some independently credible commitments? (Ebertz, 1993) In reply, we first note that it is plausible to think that the agent's input commitments reflect what she takes to have independent credibility. Surely, it would be bizarre of an agent to rely on input that she does not take to have credibility; and if that input does not go back to an earlier application of RE, it can only have the credibility as independent credibility. Second, it is plausible that a normal agent's assessment regarding independent credibility indeed tracks credibility reasonably well. It then follows that the input commitments are likely to have independent credibility, which they keep when they become part of the final RE state. To argue for this point, we need not confuse independent credibility with perceived independent credibility. The point is rather that both are correlated as a matter of fact. Admittedly, this reply presupposes that epistemic agents can to some degree judge independent credibility even if such judgements are often difficult and defeasible[17]. But this is a presupposition that all who are not sceptics concerning epistemic justification have to make (as Kelly & McGrath, 2010 underline in their critique of RE).

What we have said about independently credible commitments constrains neither their content (they may range from most general to very specific propositions) nor the firmness with which they are held. We can also allow that an agent ends up discounting the independent credibility of, say, all her specific moral convictions because she finds that she is much stronger committed to the idea that her moral views are just hopeless results of prejudice and upbringing.

Admittedly, though, our weakly foundationalist reply to the *ex-nihilo* charge leads to further questions: First, which commitments do have independent credibility, after all? And second, how much independent credibility is needed?

In our view, RE theorists need not answer the first question. RE is a theory of the general structure of justification, not an account of how specific commitments obtain independent credibility. For an account of this sort, we need substantial epistemic theories of, for example, observation, memory, intuition, testimony, introspection or analytic reasoning (and such theories may enter a RE as background theories). From the viewpoint of a RE theorist, such theories need not imply that some type of commitments (e.g. those based on observation) have always a certain degree of indepen-

[16] At this point, one might want to object that, if an input commitment survives the RE process because it has been formed in accordance with an epistemic background theory, this is again a matter of coherence, and thus not something that underwrites independent credibility. But note first that such coherence with epistemic background theories is not what was originally targeted by the objection from coherence, which targeted the relation between commitments and 'foreground' theory. In any case, the claim that epistemic background theories underwrite credibility does not imply a coherentist account of justification since such a claim is part of foundationalist epistemologies as well.

[17] See Elgin (Elgin, 1996: 101–6; 2017:64–6) on "initial tenability" for an extended discussion.

dent credibility. All that the RE theorist needs is that, in a state of RE, the epistemic theories ascribe independent credibility to some commitments.

The second question, though, is vital for RE theorists: How many commitments with which degree of independent credibility are needed in a RE state if this state is to constitute a justification for the commitments and the theory it comprises? A detailed answer can only be worked out with the help of a formal analysis (see van Cleve, 2011; 2014; Roche, 2012). The key challenge for a weak foundationalist interpretation of RE is to show that a RE state can constitute justification even if no commitment has a degree of independent credibility that suffices for knowledge (or an alternative type of epistemic target state). In short, one needs to show that coherence – possibly in combination with doing justice to epistemic goals – can boost credibility to the level needed for justification even if no commitment is sufficiently justified independent of its coherence.[18] In what follows, we assume that this challenge can be met.

2.6 Target state: RE is not truth-conducive

A particularly common concern about RE is that it does not meet the key requirement for theories of epistemic justification: it is not (sufficiently) truth-conducive. This is not only a stock objection against coherentism, but remains relevant, even if RE is decoupled from coherentism, as proposed in the previous section. In its most direct form, the worry is that being coherent (see, e.g. Olsson, 2005), respecting input commitments (e.g. Little, 1984) and doing justice to epistemic goals do not provide sufficient reason to think that a position in RE (or at least a large part of it) is likely true (e.g. Kappel, 2006). Further, even if one is willing to concede that some aspects of RE, for example, consistency and mutual support, are truth-conducive, one might worry that they conflict with other aspects of RE that are not taken to be truth-conducive, for example, theoretical virtues such as simplicity. Trade-offs between the various desiderata may then compromise RE's truth-conduciveness. Worries of this kind may be fuelled by the fact that prominent defenders of RE such as Rawls (1975; 1980) and Elgin (1996; 2017) have distanced RE from truth.

The objection hinges on the expectation that RE (in fact, epistemic justification in general) is truth-conducive. This expectation, in turn, is typically motivated by the assumption that knowledge is the ultimate epistemic goal (e.g. McGrath, 2019). Some proponents of RE share this expectation. Tersman, for example, holds that RE is "a theory about when moral beliefs are justified relative to the aim of uncovering the truth about moral issues" (2018:2; see also 1993:ch. 5.2). Elgin (1996; 2017), by contrast, argues that understanding rather than knowledge is the pivotal cognitive objective and that truth plays therefore only a limited role.

Accordingly, there are two main strategies for dealing with the objection. One option is to argue that RE is sufficiently truth-conducive because the incriminated

[18] Note that this line of argument is only meant to show that RE can be defended against the coherentism objections under discussion. It alone cannot establish that RE is superior to, or at least on a par with, moderate foundationalism. Comparative evaluation of RE and rival theories of justification is beyond the scope of this paper, which only aims to show that there is a conception of RE that can be defended against all prominent objections.

aspects of RE that do not seem truth-conducive are in fact so or at least suitably related to truth. In this vein, one may try to argue that simplicity is connected to truth because simpler theories are easier to test for truth (e.g. Douglas, 2013) and that this gives simpler theories an advantage related to truth. The alternative option, which we adopt here, is to hold that that epistemic justification is not exclusively related to truth. As we see it, RE is not an account of the kind of justification that is required for knowledge, but rather of justification tailored to the epistemic objective of understanding a subject matter by means of a theory (Baumberger & Brun, 2017, 2021; Elgin, 1996, 2017). Considerations familiar from philosophy of science show that, given this objective, justification cannot be interpreted exclusively in terms of truth-conduciveness. Epistemic evaluation of theories is rather a matter of a plurality of epistemic virtues or goals (as discussed in Sect. 2.4), some of which are not instrumental to the pursuit of truth, but rather instantiate independent, intrinsic values (see Kuhn, 1977; Hempel, 2000). If there are several independent epistemic goals, they can, and often will, pull in different directions and therefore make trade-offs inevitable, for example, when a gain in precision can only be had at the expense of scope. Likewise, in moral theorizing, considerations of applicability may favour adopting a principle that is not as accurate as one might hope. According to our view, then, trade-offs are not problematic, but rather something we should expect anyway.

2.7 Target state: RE fails to lead to a unique justified state

Starting with Hare and Singer's reviews of *A Theory of Justice*, RE's detractors have argued that the process of equilibration can lead to different, even incompatible RE states. This diagnosis, in turn, was the basis for the charges that RE is unacceptably relativist (Singer, 1974:501) or subjectivist (Hare, 1973:145–6), that RE fails to ensure objectivity (e.g. Singer, 1974:494; de Maagt, 2017) and even the worry that everything and hence nothing can be justified with RE (Haslett, 1987). The basic diagnosis of non-uniqueness can be bolstered with reference to RE classics. Goodman consistently advocated what he called "radical relativism under rigorous restraints" (1978:x) and consequently underlined that different sets of principles may be justified using RE (1983:63); Rawls admitted the possibility of plural RE states (e.g. 1975; cf. 1999:44); and Elgin has emphasized that adopting RE leads to pluralism (1996:135).[19]

We can address the objection from non-uniqueness by considering how agents can end up with different RE states about the same topic. A first possibility is that different agents pursue different epistemic goals, that they differ significantly in the weights they assign to their goals or that they draw on different background theories. But in such cases, the result that they reach different RE states hardly constitutes an objection against RE. Clearly, different epistemic goals or different trade-offs between goals can favour different theories. It is also reasonable to expect that several agents may not reach the same position in, say, moral questions if they start with markedly

[19] We leave aside versions of the objection which rest on the assumption, already discussed and criticized in Sect. 2.4 and 2.5, that RE requires only to reach a consistent position in which its elements support each other (e.g. Haslett, 1987 and Little, 1984 hold).

different background theories about rational decisions or moral epistemology. It is too much to require that an account of justification secures agreement in such cases.

A second possibility seems more worrisome. Different agents could arrive at different RE states just because they hold different input commitments (Hare, 1973; Singer, 1974). Either, such differences between the agents may not be resolved, or they may even be amplified during the equilibration process. Two strategies are available to answer this version of the objection. The first strategy is to argue that the objection overrates the possibility of reaching different RE states. As noted in Sect. 2.4, many opponents seem to underestimate the revisionary effects of RE and therefore overstate the dependence of the resulting state on the input commitments (see Freivogel, 2023 for a formal analysis). In the literature, two further arguments have been proposed to show that there may be less room for diverging reflective equilibria than it might seem. For one thing, learning that another agent has reached a different RE state is a reason to consider this agent's commitments (and her choice of candidate principles and steps of adjustment; see Scanlon, 2003:152–3). This, in turn, can generate some pressure to revise one's own commitments in a direction that makes it more likely that the agents reach the same RE state (Elgin, 1996:111–8; 2005; Tersman, 2018). For another thing, Tersman (1993:ch. 5.3) has given an extended argument that the requirement of coherence built into RE makes it unlikely and, at some level of coherence, virtually impossible that *p* is justified for one agent while *not-p* is justified for another agent. Both arguments cannot be discussed adequately here because they rest on assumptions (about opinion dynamics and holism, respectively) that would require an in-depth analysis.

Another, independent, strategy is to argue that at least some versions of the objection rest on expectations that are too high. By itself, the possibility of reaching different RE states from different input commitments is not a convincing objection to an account of epistemic justification (see, e.g. Kelly & McGrath, 2010). Since justification does not entail truth, it does not follow that two incompatible propositions cannot both be justified. According to most epistemologies, including realist views that strongly prioritize truth, different agents can be justified in having different commitments (e.g. because they have access to different evidence) and in adhering to different principles (e.g. because they interpret the same evidence differently). Non-uniqueness in this sense is no basis for the mentioned charges of relativism, subjectivism and the like (Tersman, 1993:21–2, 99, 103).

The third way in which agents may reach different RE states is this: the agents rely on the same input commitments, background theories and epistemic goals, but still reach rival RE states because they carry out the process of equilibration in different ways. This may happen for two types of reasons. First, epistemic agents have only limited cognitive resources. Therefore they may have to decide, for example, in which order to address conflicts without being able to explore all consequences of all possible alternative orders (Bonevac, 2004), and depending on the decision they take, they can reach different RE states (path-dependence). Second, agents can be confronted with underdetermination, for example, when they face a choice between positions that are equally good ("tie") or incommensurable (Elgin, 1996:134–7; Brun, 2022). Both cognitive limitations and underdetermination can create some leeway

already for one single agent. Hence, we have to reckon with the possibility that the agent herself, or someone else, could have reached an alternative RE state.

Proponents of RE might hope that the freedoms agents have in the equilibration process arise only because RE lacks specification so far; and that a more specific description of RE could determine each step in the process clearly. However, it is illusory to hope that more specific characterizations of RE processes eliminate all the non-deterministic aspects mentioned. As far as agents with limited resources and capabilities are concerned, a sensible specification of the RE process (see Sect. 2.8) will likely have to allow some arbitrary choices. For instance, realistic agents cannot survey all of their commitments on a topic; they rather have to start with some selection of commitments, and there is likely no way to determine that starting point uniquely. Or in the case of path-dependency, it is true that ideal agents could explore all orders of dealing with conflicts and identify a path that leads to a maximally good final state. But in most situations, agents with limited resources cannot do this. Nor can we expect epistemologists to come up with rules that specify in general terms which order will lead to the best outcome. As far as underdetermination is concerned, the problem is that RE involves trade-offs between several desiderata and that ties will therefore be possible. This problem cannot be solved by further specifying the desiderata.

The question then is whether non-uniqueness due to cognitive limitations or to underdetermination is a significant problem for RE. In our view, this is not the case; rather, we should give up the expectation that RE processes always lead to unique outcomes. In the case of non-uniqueness due to cognitive limitations, the reason is simply that such non-uniqueness is a most general problem that arises for every theory of justification. It just seems impossible to provide simple rules that uniquely determine what limited agents should do to obtain justification. In the case of non-uniqueness due to underdetermination, our epistemic standards fail to completely determine an epistemic choice between several positions. But such underdetermination is clearly compatible with justification if it is the result of a plurality of epistemic standards which can conflict and thus lead to a tie (see Tulodziecki, 2012). Consider an example from physics. There are several markedly different attempts to solve the so-called measurement problem of quantum mechanics, for example, Bohmian mechanics and the so-called GRW theory. A definite decision between these approaches has proven difficult because each approach surpasses others regarding some standards while falling behind the rivals on others; and it is plausible to think that we are here running into an underdetermination problem (Egg & Saatsi, 2021). It would be weird to try to avoid this problem by coming up with a new theory of justification. Indeed, alternative theories of justification leave the possibility of underdetermination, too.

Of course, if agents have cognitive limitations or face problems of underdetermination, the question arises how to take this into account when appraising the position the agent has reached. Should we say, for example, that known underdetermination lowers the degree of justification? If our analysis is right, such questions arise independently of whether one endorses RE as an account of justification, and therefore they are no basis for objections specifically targeting RE. So even though RE theorists should address these questions, they fall outside the scope of this paper.

All in all, the result is that, to the extent to which it affects RE, non-uniqueness is not a problem for RE, because justification must generally admit such non-uniqueness, contrary to what defenders of the objection seem to have assumed.

While this section has focused on the complaint that more than one RE state can be reached, there is also the opposite worry that no RE states can be reached because RE is too demanding.

2.8 General objection: RE is too demanding for real agents

The objection that RE is too demanding comes in different versions, which all rest on the expectation that RE should be something that can be achieved not only by idealized but also by real agents with limited capacities or bounded rationality. Whereas objections in the literature concentrate on the state of RE, one can also focus on the RE process.

Being in a RE *state* requires a lot. First of all, the commitments and the theory need to be consistent and support each other as demanded by coherence. Additional requirements discussed in the preceding sections include support by background theories, doing justice to epistemic goals, respecting input commitments, and independent credibility of some resulting commitments. Since not even consistency is easily attained by real agents, the conclusion that real agents cannot achieve RE states seems quite plausible (see, e.g. Beauchamp & Childress, 2013:410; DePaul, 2011:ciii; Raz, 1982). The difficulty to achieve a RE state is particularly clear for conceptions of RE which emphasize the holistic nature of RE states by suggesting that all relevant background theories should be included and potentially also be subject to revisions during the process (e.g. DePaul, 2011:lxxxix-xc; Elgin, 2017:ch. 4). The objection becomes most pressing (as argued in Arras, 2007) if RE states are subject to an optimality condition. For instance, some RE theorists hold that RE states require maximum coherence (Brink, 1989:131) or maximum independent credibility (Scheffler, 1954), or must be as good as alternative candidate RE states. Checking whether a given position fares at least as well as potential alternatives with respect to the criteria for RE states seems out of reach for beings with limited cognitive resources, because alternative candidates for RE states are not only too numerous but also not readily available. It is even difficult to have a clear overview of what the space of the relevant alternatives is. Defenders of optimality constraints have of course been aware of such problems and weakened the conditions. Rawls (1999:43), for example, was quick to add that for practical reasons we may confine attention to plausible alternative theories that we know of or that occur to us. And Elgin emphasizes that a position in RE need only be "as good as any available alternative" (2017:87–8). Nonetheless, the worry persists that agents with limited capacities may not be able to reach RE states.

There are several strategies to respond to the objection. One option is to insist that, for some specific subject matters, RE states have in fact been achieved. Arguably, logicians and mathematicians have reached reflective equilibria at least for the

elementary parts of logics and arithmetic. This would show that the conditions for RE states can in principle be met.[20]

Second, for some subject matters, one can blame the problem of demandingness not on RE but on the subject matter. Maybe, ethics is just so hard that it is premature to expect a comprehensive account of ethics in reflective equilibrium presently (e.g. DePaul, 2013:4473).

A third strategy attacks the expectation that RE states should be realistically reachable and suggests that RE states should rather be interpreted as an ideal (Arras, 2007; Knight, 2017). It is then no longer essential that RE states can be reached; still, making progress with justification can be explained as getting closer to the ideal (DePaul, 2011:ciii; Scanlon, 2014:77; see also Tersman, 1993:45–6). This reply presupposes that various epistemic states can be compared to each other with respect to their distance from the ideal. Such a comparison makes indeed sense insofar as all desiderata that characterize RE states (coherence, epistemic virtues etc.; except consistency) can be realized to a greater or lesser extent.

Interpreting RE states as an ideal dovetails with the stance (taken in Sect. 2.6) that RE is an account of the justification needed for understanding a subject matter by means of a theory. For understanding, in contrast to outright theory acceptance, maximality requirements are not plausible since even sub-optimal theories, for example, from past science, may provide some understanding of their subject matter although they do not realize the ideal fully or better-justified alternatives are available (see Baumberger & Brun, 2017 for further discussion).[21]

If we now turn to the *process* of equilibration, one worry is that each individual step of the process is too demanding for agents with limited capacities. At the very beginning, the agent has to come up with a theory that promises to account for the initial commitments, and at every subsequent step, the agent needs to decide on how to adjust the commitments or the theory. However, surveying comprehensively candidates for theories or possible adjustments seems often out of reach for human agents.

One strategy to defuse this problem is to insist on the piece-meal character of the process by, for example, requiring that all but the first steps change at most one or only a few commitments or elements of the theory.[22] This reduces the problem, because it is clearly more feasible to select just a few commitments or principles for adjustment. However, there is a further worry, which is not necessarily attenuated and possibly even aggravated by working with piece-meal adjustments. There is a certain likelihood that, at a given step, the agent has a choice between several equally viable adjustments. And this raises the question of whether the agent needs to branch the process and check out for each candidate adjustment how the process could be continued. But then a large number of possible sequences of adjustments

[20] One might worry that such examples are plausible only as long as we consider a relatively narrow equilibrium; that is, an equilibrium which involves only a few background theories. See, e.g. Daniels, 1980 on the role of theories of meaning as background theories for logic.

[21] For a discussion of objections from demandingness in the context of justification required for knowledge, see McGrath, 2019:ch. 2.2.

[22] Goodman's original description of RE may be read as referring to such piece-meal adjustments: "*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." (Goodman, 1983:64; our italics).

may become available and a real agent would often not be able to get an overview of all of them, not to mention the efforts needed to work through all of them. Presently, the question of whether a piece-meal strategy makes RE processes more manageable for real agents must be left unanswered since detailed investigations and even precise descriptions of RE processes have hardly been attempted so far. We will come back to this lacuna in Sect. 4.

Another strategy to counter the process-related demandingness objection is to insist that descriptions of the equilibration process should, again, be interpreted as an ideal we can try to approximate when trying to make progress with our views on a subject matter. What this amounts to depends on how we think of making progress in a RE process. Do we think of the steps in the RE process as making progress just in case they bring us closer to a RE state? Or should we think of certain steps as inherently progress-making? Or maybe both? Here, we reach conceptual issues we think are fundamental for RE theory, yet severely underexplored in the literature (see Sect. 4). As long as RE theorists have not addressed such issues in more depth, it is impossible to appraise the objection from demandingness.

All in all, our discussion of objections from demandingness has a double upshot. The idea of a RE should be understood as invoking an ideal that can be approximated with the resources of real agents, and the process of equilibration as well as its relation to RE states need to be investigated much more thoroughly by RE theorists.

3 A proposal for a re-conception of RE

In our discussion of the most important objections against RE, we have indicated how we think RE theorists should deal with them, and we have seen that our favoured replies have repercussions on the conception of RE. But are our replies consistent and if so, what do they imply for our conception of RE? This section aims to answer these questions.

We start with a short overview of the points we have made. The *first* objection (RE is an uninformative account of justification) has been taken as a challenge to come up with a more informative specification of RE. In our reaction to the *second* objection (RE is tied to intuitions), we have separated RE from intuitions and suggested that input commitments of various sources must be examined with the help of epistemological theories. The consequence for RE is that RE leaves open where the input commitments come from, but subjects them to critical scrutiny using epistemological background theories. The *third* objection (garbage in, garbage out) has to some extent been rebutted by questioning the expectations behind it: RE does not need to lead to a RE state for every input. But we have also argued that a substantial response is needed, a response we have also used to rebut the *fourth* objection (RE is too conservative). In our view, RE has more revisionary potential than many have recognized. This is so if RE is wide and requires the consideration of background theories. Further, epistemic goals or theoretical virtues make RE more powerful. Our response to the *fifth* objection (RE is tied to coherentism) has essentially been to introduce independent credibility as an additional source of justification. Thus, a position and its elements are not justified only because they are coherent (also with

background theories) and satisfy the epistemic goals. What is additionally needed is independent credibility of the commitments in the final state. This requirement separates RE from coherentism and associates it with weak foundationalism. The *sixth* objection (RE is not sufficiently truth-conducive) can be countered either by insisting that the emerging conception of RE leads to the truth reasonably well or by associating RE to understanding rather than knowledge. In this paper, we have taken the second route. To the *seventh* objection (RE does not ensure that a unique justified state is reached), we have replied by analysing the ways in which a plurality of final RE states can arise. According to our results, it is trivial and unavoidable that different epistemic goals can lead to different RE states. Likewise, there is no problem if different input commitments lead to different RE states, since justification is relative to an epistemic agent. Finally, RE may lead to differing positions because realistic agents cannot explore all epistemic options or because certain choices are underdetermined. But neither possibility constitutes a problem since, under the conditions of limited agents or underdetermination, we cannot expect that justification is attributed only to one single state. The *eighth* and last objection (RE is too demanding) was countered by interpreting RE states as ideals and by identifying a need for more research on RE processes.

We believe that this leaves us with a consistent conception of RE that usefully spells out the pre-conception we have started with. To show this, let us go through the conception in a more systematic manner. We begin with static aspects of RE; that is, aspects that do not refer to a process.

First, there is the *aim* of RE. From the outset, we have interpreted RE as an account of epistemic justification. We have then suggested that justification by RE is tailored to understanding rather than to knowledge and argued that this justification need not be truth-conducive.

It is also important to note that RE specifies an ideal. In practice, we can only approximate this ideal, both regarding RE states and the process.

Consider now the *scope* of RE. In our view, RE is not restricted to specific topics or questions. As a general account of justification, it can only specify the relevant elements and the general structure of justification. If RE is applied to specific fields, for example to ethics, it has to be coupled with more specific epistemological theories about the sources of commitments. Such theories then play the role of background theories.

Regarding the *components* of RE, we have argued that, in addition to commitments and theories, two further elements must be included. Firstly, epistemic goals are crucial to RE, especially theoretical virtues such as simplicity. They are the ultimate driving force behind the dynamics of RE because they require the epistemic agent to develop a *systematic* (and thus more than a trivial) theory of the subject matter. Secondly, RE has to be wide and thus include background theories; that is, theories which are expected to provide support for the agent's position. As said before, background theories are in particular expected to include epistemic theories which permit a critical evaluation of input commitments.

Finally, there is what may be called the *axiology* of RE. This comprises the desiderata which decide whether a position is justified, or more justified than alternatives. Internal and external coherence of the resulting position are certainly important, but

we have argued that they do not suffice. Internal and external coherence ensure that the agent's commitments are consistent and can be inferred from a theory which is appropriately supported by background theories. However, the ideal of a RE state additionally requires that the resulting commitments adequately respect input commitments, that the resulting theory does justice to the relevant epistemic goals and that some commitments have some credibility independent of their coherence with other commitments. So, all in all, RE as a target state is a matter of meeting four conditions: coherence, doing justice to epistemic goals, respecting input commitments and independent credibility.

We turn now to the *dynamics* of RE; that is, the process of mutual adjustment of commitments and theory. It all begins with an initial state constituted by the agent's initial commitments. The agent then carries out an equilibration process until she reaches a final state, which ideally is a RE state. At the stages between the initial and the final state, the agent makes adjustments to the commitments or to the theory. At all these stages, the epistemic agent can adopt new commitments, either as a result of deriving them from the theory she currently entertains or for other reasons, for example, because she acquires new evidence or considers new problems. Commitments of the latter type are counted among the input commitments. For a further specification of the process of equilibration, we need to incorporate two key features. Firstly, the agent's position is extensively revisable. This means that each commitment and each element of the theory can in principle be revised. Secondly, the process is driven by the goal of reaching a RE state as characterized by the axiology of RE. We do not assume that applying RE always leads to such a RE state, but since RE defines an ideal, the process remains useful to the extent that it brings agents closer to a RE state.

Interestingly, this conception of RE contains two interpretations of the metaphor of an equilibrium. On the one hand, an agent's position must be in equilibrium insofar as her commitments must be in agreement with the theory. On the other hand, a position in reflective equilibrium must be stable between the pull of two forces, the 'conservative' pull of input commitments and the 'progressive' pull of the epistemic goals.

4 Outlook: pressing issues for defenders of RE

We think that the conception of RE sketched in Sect. 3 has a good chance of overcoming all objections discussed in Sect. 2. But we also think that our discussion of objections has highlighted some issues to which proponents of RE should devote more attention. In our view, the most important problems appeared in Sect. 2.8 (about the demandingness objection) and concern the process of equilibration and its relation to RE as a target state. If we want to address this point more thoroughly, we face a deeper conceptual issue that has not yet been made explicit. Why should RE include both, a static and a dynamic, component? Is one of them fundamental? Two straightforward positions suggest themselves. One gives priority to the RE state and holds that we can describe the state of RE and its good-making features without referring to the process. The process is then interpreted in purely instrumental terms, as a means to get to a state that is optimal in view of the good-making features. In contrast

to such a "consequentialist" outlook, a "proceduralist" view holds that the process has priority and that justification is simply achieved by a state at which the process finishes because that follows from the rules that define the procedure. There is also space for options that avoid claiming the priority of either the state or the process and consequently reject both consequentialism and proceduralism. Although some considerations can be found in Elgin, 1996 and Baumberger & Brun, 2021, a much more thorough analysis of consequentialist vs. proceduralist interpretations of RE and of the function of the RE process in general is still a desideratum of RE theory.

A second important question that our discussion has left open is to what extent commitments must carry independent credibility to avoid worries about coherentism. As we argued, work in formal epistemology, addressing this question in a setting of RE, will be needed for an answer.

More specific questions about the RE process require additional research as well. First of all, exploring whether and how the process of equilibration could be specified more exactly would be most important. It is surprising how little RE theorists have so far discussed the details of the process. And further questions have hardly been examined, for example: Can general, subject-independent, rules for making adjustments to commitments and theories be given? Can such rules be specified in a formal way, possibly relying on existing theories of belief dynamics? Is it possible to specify the process in a way that ensures that each RE state can be reached by carrying out the process?[23] Answering such questions requires, we believe, not only conceptual efforts but also case studies and formal models – hence, much work is still to be done in RE theory.[24] Hopefully, this paper has at least shown that such efforts are not in vain. There is, after all, a conception of RE that can meet the most pressing objections, or so we have argued.

Acknowledgements Earlier versions were presented in Berne, Karlsruhe and Munich. For helpful discussions and comments, we thank the audiences and, in particular, Gregor Betz, Thomas Schmidt and Folke Tersman, as well as the anonymous reviewers. Research for this paper is part of the project *How far does Reflective Equilibrium Take Us? Investigating the Power of a Philosophical Method* (SNSF grant 182854 and German Research Foundation grant 412679086).

Funding Open access funding provided by University of Bern

[23] For an initial discussion of the expectation that general yet precise rules for the equilibration process could be given see Walden, 2013 and Baumberger & Brun, 2021.

[24] For formal models of RE processes, see Thagard, 2000; Yilmaz et al., 2017; Beisbart et al., 2021. On relying on existing theories of belief dynamics, see Bonevac, 2004 and Freivogel, 2021. On addressing objections to RE with the help of a formal model, see Freivogel, 2023.

References

Appiah, K. A. (2008). *Experiments in Ethics*. Harvard University Press.

Arras, J. D. (2007). The way we reason now. Reflective equilibrium in Bioethics. In B., Steinbock (Ed.), *The Oxford Handbook of Bioethics* (pp. 46–71). Oxford University Press.

Baumberger, C., & Brun, G. (2017). Dimensions of Objectual understanding. In S. R. Grimm, C. Baumberger, & S. Ammon (Eds.), *Explaining understanding. New perspectives from Epistemology and Philosophy of Science* (pp. 165–189). Routledge.

Baumberger, C., & Brun, G. (2021). Reflective equilibrium and understanding. *Synthese*, *198*, 7923–7947.

Beauchamp, T. L., & Childress, J. F. (2013). *Principles of Biomedical Ethics* (7th ed.). Oxford University Press.

Beisbart, C., Betz, G., & Brun, G. (2021). Making reflective Equilibrium Precise. A formal model. *Ergo*, *8/15*, 441–472.

Bonevac, D. (2004). Reflection without equilibrium. *Journal of Philosophy*, *101*, 363–388.

BonJour, L. (1985). *The structure of empirical knowledge*. Harvard University Press.

Brandt, R. B. (1979). *A theory of the good and the right*. Clarendon Press.

Brandt, R. B. (1985). The Concept of Rational Belief. *The Monist*, *68*, 3–23.

Brink, D. O. (1989). *Moral Realism and the foundations of Ethics*. Cambridge University Press.

Brun, G. (2014). Reflective equilibrium without intuitions? *Ethical Theory and Moral Practice*, *17*, 237–252.

Brun, G. (2020). Conceptual re-engineering: From explication to reflective equilibrium. *Synthese*, *197*, 925–954.

Brun, G. (2022). Re-engineering contested concepts. A reflective Equilibrium Approach. *Synthese*, *200*, 168.

Burkard, A. (2012). *Intuitionen in Der Ethik*. Mentis.

Cath, Y. (2016). Reflective equilibrium. In H. Cappelen, T. S. Gendler, & J. Hawthorne (Eds.), *The Oxford Handbook of Philosophical Methodology* (pp. 213–230). Oxford University Press.

Cohen, L. J. (1981). Can Human rationality be experimentally demonstrated? *The Behavioral and Brain Sciences*, *4*, 317–331.

Daniels, N. (1979). Wide Reflective Equilibrium and Theory Acceptance in Ethics. *The Journal of Philosophy* 76, 256–82. Reprinted in Daniels 1996:21–46.

Daniels, N. (1980). On Some Methods of Ethics and Linguistics. *Philosophical Studies* 37, 21–36. Reprinted in Daniels 1996:66–80.

Daniels, N. (1996). *Justice and Justification. Reflective equilibrium in theory and practice*. Cambridge University Press.

Daniels, N. (2018). Reflective Equilibrium. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy*. https://plato.stanford.edu/archives/spr2018/entries/reflective-equilibrium/.

de Maagt, S. (2017). Reflective equilibrium and Moral Objectivity. *Inquiry: A Journal of Medical Care Organization, Provision and Financing*, *60*, 443–465.

De Regt, H. W. (2017). *Understanding Scientific understanding*. New York.

DePaul, M. R. (1998). Why bother with reflective equilibrium? In M. R. DePaul, & W. Ramsey (Eds.), *Rethinking intuition. The psychology of intuition and its Role in Philosophical Inquiry* (pp. 293–309). Rowman and Littlefield.

DePaul, M. R. (2006). Intuitions in Moral Inquiry. In D. Copp (Ed.), *The Oxford handbook of ethical theory* (pp. 595–623). Oxford University Press.

DePaul, M. R. (2011). Methodological issues. Reflective equilibrium. In C. Miller (Ed.), *The Continuum Companion to Ethics* (pp. lxxv–cv). Continuum.

DePaul, M. R. (2013). Reflective equilibrium. In H. Lafollette (Ed.), *The International Encyclopedia of Ethics* (pp. 4466–4475). Wiley-Blackwell.

de Sousa, R. (2010). Here's how I feel. Don't trust your feelings! In S. Roeser, (Ed.), *Emotions and risky technologies* (pp. 17–35). Springer

Douglas, H. (2013). The value of cognitive values. *Philosophy of Science*, *80*, 796–806.

Ebertz, R. P. (1993). Is reflective equilibrium a Coherentist Model? *Canadian Journal of Philosophy*, *23*, 193–214.

Egg, M., & Saatsi, J. (2021). Scientific realism and underdetermination in Quantum Theory. *Philosophy Compass*, *16*, e12773.

Elgin, C. Z. (1983). *With reference to reference*. Hackett.

Elgin, C. Z. (1996). *Considered Judgment*. Princeton University Press.
Elgin, C. Z. (2005). Pragmatism, historicism, and/or reflective equilibrium. *Philosophy of Education*, *61*, 57–59.
Elgin, C. Z. (2014). Non-foundationalist Epistemology. Holism, Coherence, and Tenability. Reply to Van Cleve. In M. Steup, J. Turri, & E. Sosa (Eds.), *Contemporary debates in Epistemology* (2nd ed., pp. 244–255). 267–71). Wiley.
Elgin, C. Z. (2017). *True enough*. MIT Press.
Foley, R. (1993). *Working without a net. A study of egocentric epistemology*. Oxford University Press.
Freivogel, A. (2021). Modelling reflective equilibrium with belief Revision Theory. In M. Blicha, & I. Sedlár (Eds.), *The Logica Yearbook 2020* (pp. 65–80). College Publications.
Freivogel, A. (2023). *Does Reflective Equilibrium Help Us Converge? Synthese* 202, 171.
Fumerton, R. (2009). The Problem of the Criterion. In J. G (Ed.), *The Oxford Handbook of Skepticism* (pp. 34–52). Oxford University Press.
Goodman, N. (1952). Sense and Certainty. *The Philosophical Review* 61, 160–67. Reprinted in (1972). *Problems and Projects* (pp. 60–8). Indianapolis/New York: Bobbs-Merrill.
Goodman, N. (1978). *Ways of Worldmaking*. Hackett.
Goodman, N. (1983). [1954]. *Fact, fiction, and Forecast* (4th ed.). Harvard University Press.
Hare, R. M. (Ed.). (1973). Rawls' Theory of Justice. *The Philosophical Quarterly* 23, 144–55, 241–52. Reprinted in N. Daniels (Ed.), (1975). *Reading Rawls. Critical Studies on Rawls'* A Theory of Justice (pp. 81–107). Oxford: Basil Blackwell.
Haslanger, S. (1999). What Knowledge Is and What It Ought to Be. Feminist Values and Normative Epistemology. *Philosophical Perspectives* 13 Epistemology, 459–80. Reprinted in (2012). *Resisting Reality. Social Construction and Social Critique* (pp. 341–64). Oxford: Oxford University Press.
Haslett, D. W. (1987). What is wrong with reflective Equilibria. *Philosophical Quarterly*, *37*, 305–311.
Hempel, C. G. (2000). [1988]. On the Cognitive Status and the Rationale of Scientific Methodology. *Selected philosophical essays* (pp. 199–228). Cambridge University Press.
Huemer, M. (2005). *Ethical intuitionism*. Palgrave Macmillan.
Kappel, K. (2006). The Meta-Justification of reflective equilibrium. *Ethical Theory and Moral Practice*, *9*, 131–147.
Keas, M. N. (2018). Systematizing the theoretical virtues. *Synthese*, *195*, 2761–2793.
Keefe, R. (2000). *Theories of vagueness*. Cambridge University Press.
Kelly, T., & McGrath, S. (2010). Is reflective equilibrium Enough? *Philosophical Perspectives*, *24*, 325–359.
Knight, C. (2017). Reflective equilibrium. In A. Blau (Ed.), *Methods in Analytical Political Theory* (pp. 46–64). Cambridge University Press.
Knight, C. (2023). Reflective Equilibrium. In E. N. Zalta & U. Nodelman (Eds.), *The Stanford Encyclopedia of Philosophy*. https://plato.stanford.edu/archives/win2023/entries/reflective-equilibrium/.
Kuhn, T. S. (1977). Objectivity, Value Judgment, and Theory Choice. *The essential tension. Selected studies in scientific tradition and change* (pp. 320–339). University of Chicago Press.
Lewis, D. (1983). *Philosophical papers. Vol I*. Oxford University Press.
Lewis, D. (1986). [1973]. *Counterfactuals* (reprint with corrections). Oxford: Blackwell.
Little, D. (1984). Reflective equilibrium and justification. *Southern Journal of Philosophy*, *22*, 373–387.
Lycan, W. G. (2011). Epistemology and the role of intuitions. In S. Bernecker, & D. Pritchard (Eds.), *The Routledge Companion to Epistemology* (pp. 813–822). Routledge
Lyons, D. (1975). Nature and Soundness of the contract and coherence arguments. In N. Daniels (Ed.), *Reading Rawls. Critical studies on Rawls' a theory of Justice* (pp. 141–167). Basil Blackwell.
McGrath, S. (2019). *Moral Knowledge*. Oxford University Press.
Mikhail, J. (2010). Rawls' Concept of reflective equilibrium and its original function in a theory of Justice. *Washington University Jurisprudence Review*, *3*, 1–30.
Mikhail, J. (2011). *Elements of Moral Cognition. Rawls' linguistic analogy and the Cognitive Science of Moral and Legal Judgement*. Cambridge University Press.
Olsson, E. J. (2005). *Against coherence. Truth, Probability, and justification*. Clarendon Press.
Paulo, N. (2020). The unreliable intuitions Objection against reflective equilibrium. *The Journal of Ethics*, *24*, 333–353.
Rawls, J. (1975). The Independence of Moral Theory. *Proceedings and Addresses of the American Philosophical Association* 48, 5–22. Reprinted in (1999) *Collected Papers* (pp. 286–302). Cambridge, MA: Harvard University Press.

Rawls, J. (1980). Kantian Constructivism in Moral Theory. *The Journal of Philosophy* 77, 515–72. Reprinted in (1999) *Collected Papers* (pp. 303–58). Cambridge, MA: Harvard University Press.
Rawls, J. (1999). *A theory of Justice. Revised edition*. Belknap Press.
Rawls, J. (2001). *Justice as Fairness. A restatement*. Harvard University Press.
Raz, J. (1982). The claims of reflective equilibrium. *Inquiry: An Interdisciplinary Journal of Philosophy, 25*, 307–330
Rechnitzer, T. (2022). *Applying reflective equilibrium. Towards the justification of a Precautionary Principle*. Springer.
Roche, W. (2012). Witness Agreement and the Truth-Conduciveness of Coherentist Justification. *The Southern Journal of Philosophy*, *50*, 151–169.
Scanlon, T. M. (2003). Rawls on Justification. In S. Freeman (Ed.), *The Cambridge Companion to Rawls* (pp. 139–167). Cambridge University Press.
Scanlon, T. M. (2014). *Being realistic about reasons*. Oxford University Press.
Scheffler, I. (1954). On justification and commitment. *Journal of Philosophy*, *51*, 180–190.
Schindler, S. (2018). *Theoretical virtues in Science. Uncovering reality through theory*. Cambridge University Press.
Siegel, H. (1992). Justification by Balance. *Philosophy and Phenomenological Research*, *52*, 27–46.
Singer, P. (1974). Sidgwick and reflective equilibrium. *Monist*, *58*, 490–517.
Singer, P. (2005). Ethics and intuitions. *The Journal of Ethics*, *9*, 331–352.
Stein, E. (1996). *Without good reason. The rationality debate in Philosophy and Cognitive Science*. Clarendon Press.
Stich, S. P. (1990). *The fragmentation of reason. Preface to a pragmatic theory of cognitive evaluation*. MIT Press.
Stich, S. P., Richard, E., & Nisbett (1980). Justification and the psychology of human reasoning. *Philosophy of Science*, *47*, 188–202.
Tersman, F. (1993). *Reflective equilibrium. An essay in Moral Epistemology*. Almqvist and Wiksell.
Tersman, F. (2008). The reliability of Moral intuitions. A challenge from Neuroscience. *Australasian Journal of Philosophy*, *86*, 389–405.
Tersman, F. (2018). Recent work on reflective equilibrium and method in Ethics. *Philosophy Compass* 13.
Thagard, P. (2000). *Coherence in Thought and Action*. MIT Press.
Timmons, M. (2013). *Moral Theory. An introduction* (2nd ed.). Rowman and Littlefield.
Tulodziecki, D. (2012). Epistemic equivalence and epistemic incapacitation. *British Journal for the Philosophy of Science*, *63*, 313–328.
van Cleve, J. (2011). Can Coherence Generate Warrant *Ex Nihilo?* Probability and the logic of concurring witnesses. *Philosophy and Phenomenological Research*, *82*, 337–380.
van Cleve, J. (2014). Why coherence is not enough. A defense of Moderate Foundationalism. Reply to Elgin. In M. Steup, J. Turri, & E. Sosa (Eds.), *Contemporary debates in Epistemology* (2nd ed., pp. 255–267, 271–273). Wiley
Walden, K. (2013). In defense of reflective equilibrium. *Philosophical Studies*, *166*, 243–256.
Weinberg, J. M., Nichols, S., & Stich, S. R. (2008). Normativity and Epistemic Intuitions. In J. Knobe, & S. Nichols (Eds.), *Experimental philosophy* (pp. 17–45). Oxford University Press.
Williamson, T. (2007). *The philosophy of philosophy*. Blackwell.
Yilmaz, L., Franco-Watkins, A., & Kroecker, T. S. (2017). Computational models of ethical Decision-Making. A coherence-driven reflective equilibrium model. *Cognitive Systems Research*, *46*, 61–74.