Ethics, Politics & Society
A Journal in Moral and Political Philosophy

REFLECTIVE EQUILIBRIUM IS ENOUGH: AGAINST THE NEED FOR PRE-SELECTING *CONSIDERED JUDGEMENTS*

TANJA RECHNITZER
Leibniz University Hannover
tanja.rechnitzer@philos.uni-hannover.de

ORCID: 0000 0002 8795 2463

MICHAEL W. SCHMIDT
Karlsruhe Institute of Technology
michael.schmidt@kit.edu

ORCID: 0000 0002 4602 1478

(Received: 28 February 2022/ Accepted: 18 October 2022)

Abstract: In this paper, we focus on one controversial element of the method of reflective equilibrium, namely Rawls' idea that the commitments that enter the justificatory procedure should be pre-selected or *filtered*: According to him, only *considered judgements* should be taken into account in moral philosophy. There are two camps of critics of this filtering process: 1) Critics of reflective equilibrium: They reject the Rawlsian filtering process as too weak and seek a more reliable one, which would actually constitute a distinct epistemic method. 2) Proponents of reflective equilibrium: They reject the Rawlsian filtering process as too exclusionary. We defend RE against its critics, arguing that the method can secure reasonable commitments without depending on a strong external filtering process. However, we side with the critical proponents of reflective equilibrium and argue that without the Rawlsian weak filtering process, RE is more plausible both as a general method as well as in the context of moral philosophy.

Keywords: reflective equilibrium; considered judgments; John Rawls; methods of practical philosophy; epistemic filter; moral justification.

Resumo: Neste artigo focamo-nos num elemento controverso do método do equilíbrio reflexivo, nomeadamente a ideia de Rawls de que os compromissos que entram no processo justificatório deveriam ser pré-seleccionados ou *filtrados*: de acordo com Rawls, apenas os *juízos bem ponderados* deveriam ser levados em conta na filosofia moral. Os críticos deste processo de filtragem distribuem-se por dois campos: 1) Os críticos do equilíbrio reflexivo rejeitam o processo de filtragem rawlsiano por ser excessivamente fraco e procuram um que seja mais fiável e que constituiria um diferente método

epistémico; 2) Os proponentes do equilíbrio reflexivo rejeitam o processo de filtragem Rawlsiano por ser excessivamente excludente. Nós defendemos o ER, em oposição aos seus críticos, argumentando que este método pode assegurar compromissos razoáveis sem estar dependente de um processo externo forte de filtragem. No entanto, alinhamo-nos com os proponentes críticos do equilíbrio reflexivo, defendendo que, sem este processo de filtragem fraco, o ER rawlsiano revela-se como mais plausível tanto como um método geral, como no contexto da filosofia moral.

Palavras-chave: Equilíbrio reflexivo; juízos bem ponderados; John Rawls; métodos da filosofia prática; filtro epistémico; justificação moral.

Introduction

Reflective equilibrium is one of the most influential methods in philosophy. According to the method, judgements and theoretical principles are justified for an epistemic agent *iff* conflicts between them are resolved in a way such that they harmonize with each other and are part of the most plausible available system of commitments.[1]

In this paper, we focus on one controversial element of the method, namely Rawls' idea that the commitments that enter this justificatory process should be pre-selected or filtered: According to him, only *considered judgements* should be taken into account in moral philosophy. These are impartial judgements that were made in a calm mood and with confidence. According to Rawls, these are the circumstances under which our moral capacity is most likely to be displayed without distortion (Rawls 1999, p. 42). When RE is discussed as a general method of justification, these Rawlsian conditions for considered judgments (CJ-conditions) are often one of the contested elements.

There are two main lines of argument against filtering commitments with those conditions. On the one hand, the CJ-conditions are seen as too weak. This line of argument is typically pursued by critics who take the CJ-conditions to be an essential part of RE, and who reject RE as a method of justification on the grounds that this filter is too weak. They argue that the method is not enough to lead to justified results, and would need to be combined with a stronger epistemic filter, which would be external to RE.

On the other hand, the CJ-conditions are rejected as too strong, that is, too exclusionary. This line of argument is also pursued by proponents of RE, who argue that as a general method, RE would be more plausible without this Rawlsian element.

We will argue that the CJ-conditions should not be seen as an essential element of reflective equilibrium in general, and that they should be abandoned even for the area of moral philosophy. While there is some appeal to the idea of ignoring commitments that are likely to be distorted and erroneous from the start, ultimately the method of RE becomes stronger if such commitments are assessed as part of its process of analysis, evaluation, and adjustments.

[1] See Rechnitzer (2022) and Schmidt (2022) for more detailed accounts of our respective understandings of RE.

After clarifying what reflective equilibrium and considered judgments mean in the Rawlsian context (Section 1), we will reject the criticism that the CJ-conditions are too weak (Section 2). Critics that argue that reflective equilibrium is undermined as a method of justification because it would need to be combined with a strong epistemic filter overlook the holistic and social aspects of the method. We argue that they overly focus on the starting conditions and do not take the process of adjustments seriously enough. Consequently, reflective equilibrium does not depend on another epistemic method in the form of a stronger filter.

Having rejected the criticism that the CJ-conditions are too weak, we then turn to the question whether they —or a similar *weak epistemic filter*—should be seen as an essential element of reflective equilibrium. We argue that a filter for a specific inquiry would need to be justified by using the method on a meta-level without a filter. This is indeed possible but excludes it as an essential element of reflective equilibrium as a general method (Section 3).

This leads us to the question whether the use of the CJ-conditions can be defended for the area of normative ethics and political philosophy (Section 4). Ultimately, we reject this based on three main reasons: First, by filtering out moral judgments which are partial or for which we lack confidence, we are too exclusive and might lose epistemically valuable commitments. Second, making a preselection of the commitments to be considered in reflective equilibrium could undermine its critical and revisionary power. Third, the considerations behind the CJ-conditions can be integrated into the process of analysis, evaluation, and adjustments, making a previous filtering unnecessary. We thus conclude that the CJ-conditions should be rejected as a mechanism for filtering commitments before entering an RE process. Instead, the commitments that should be regarded as *considered* judgments are those that underwent a RE process and result as part of an RE state.

1. Rawlsian Reflective Equilibrium and the Conditions for Considered Judgements

Very roughly, the idea of RE is that all our commitments concerning a subject matter have to be justified through a process of analysis, evaluation, and mutual adjustments which aims to bring our commitments into the best agreement. This includes commitments from different levels of generality, like judgments on particular cases as well as theoretical considerations. That is, RE aims at a state in which our theories (or principles, models) can account best for our commitments (e.g., our judgments, beliefs, and convictions). Additionally, in this state we reflectively grasp the web of our commitments, including their inferential relations, and can explain why we are committed to the implications of our theories. One can distinguish between the RE *process of analysis, evaluation, and adjustment* and the *state* of having reached an epistemic position that is in RE; we can then say that the RE method aims at reaching the state through the process.

RE is often called a *Rawlsian method*, and both critics and proponents of the method often identify Rawls' specific conception with RE in general. Rawls himself sees RE as a universal method of justification, which is not restricted to moral philosophy. He sees it applied throughout the history of philosophy and refers to thinkers as different as Socrates, Aristotle, Henry Sidgwick, John Stuart Mill, Nelson Goodman, Willard Van Orman Quine or Morton White (Rawls, 1999, p. 18, p. 45, p. 108, p. 507). However, it makes sense to distinguish between the general idea of RE, and the specific conception that Rawls develops for moral and political philosophy. This also allows us to critically discuss elements of the Rawlsian conception—like the CJ-conditions—without conflating criticism of these elements with criticism of RE in general. In this paper, we are interested in how one can make RE in general as strong as possible. For this, however, it is central to discuss Rawls' element of filtering for considered judgments, as an important part of the discussion around RE focuses on these conditions, asking whether they are plausible as an element of RE in general or at least for the area of ethics. Let us thus now take a closer look at the Rawlsian conception, and in particular, the CJ-conditions.

According to Rawls, theories in moral philosophy—including their assumptions and models, like the original position—are internally justified for us *iff* they fit best with our coherently systematized moral commitments and thus are part of the most plausible system of commitments.[2] If our reflection succeeds in discovering or creating such a harmonious system, which consists of commitments on all levels of abstraction, we have reached a reflective equilibrium (Rawls, 1999, pp. 15–19). Rawls provides some basic outline of the way by which we should try to reach this state of reflective equilibrium, and thus, how to justify commitments. We should find out which systems of commitments would emerge if we try to systematize our moral commitments and resolve existing conflicts between them. For this purpose we go "back and forth" (Rawls, 1999, p. 18): We explore by way of trial the consequences of adjusting a moral commitment to a promising theory and its enclosed principles, versus the alternative of retaining the moral judgement by discarding the conflicting theory, considering another theory or modifying it. In this process, no commitment is in principle immune from revision, thus implying some kind of fallibilism. Due to the many possible adjustments that have some plausibility, a variety of corresponding potential systems of commitments emerges. Moreover, by going through the possible adjustments, the inferential connections between our commitments become apparent. We have acquired the ability to grasp the inferential structure of our commitments and the respective weight we assign to them. This makes it possible to identify the adjustments and the corresponding candidate system of commitments that we reflectively judge as the most plausible (Rawls, 1999, 17ff., pp. 40–46).

[2] Rawls does not make a sharp distinction between *judgment*, *belief*, or *conviction*, and is using these terms interchangeably. For systematic reasons, we adopt the more recent use of *commitment* (e.g. Rawls, 1999, 17f., p. 216, p. 220, 280f., p. 392, p. 448, 570f.; Elgin, 1996, 2017; Brun, 2013).

With respect to the method of reflective equilibrium for moral philosophy, Rawls highlights two conditions; one is concerned with inclusion, the other with exclusion:

1. Inclusion: While pursuing reflective equilibrium, one has to take into account all relevant commitments, philosophical arguments, and relevant background theories. Otherwise, the result will be too narrow-minded and inherently conservative. A wide reflective equilibrium, that results if all relevant commitments and arguments have successfully been taken into account, can have considerable revisionary force. Only a wide reflective equilibrium provides justification according to Rawls (Daniels, 1979; Rawls, 1999, p. 43; 1974).

2. Exclusion: While pursuing reflective equilibrium, one should only take into account commitments of a specific form, which Rawls calls "considered judgements" (Daniels, 1979, p. 258; Rawls, 1999, xviii, xix, 17ff., pp. 40–46).

With considered judgments, Rawls (1999, p. 42) refers to judgments that are made under "conditions favorable to the exercise of the sense of justice, and therefore in circumstance where the more common excuses and explanations for making a mistake do not obtain". As conditions for considered judgments (CJ-conditions), he names that the person making the judgment

a) wishes to arrive at true or correct judgements in the first place;

b) is clear-headed and calm when making the judgement;

c) is not threatened with any disadvantages or tempted by advantages as a result of the judgement, i.e. one is impartial (in a very weak sense);

d) has a certain confidence in the judgement and does not make it hesitantly, or can keep the judgement constant over a certain period of time (Rawls, 1999, pp. 42–43).

Rawls expands on this in *Justice as Fairness. A Restatement*:

> Considered judgments are those given when conditions are favorable to the exercise of our powers of reason and sense of justice: that is, under conditions where we seem to have the ability, the opportunity, and the desire to make a sound judgment; or at least we have no apparent interest in not doing so, the more familiar temptations being absent. Some judgments we view as fixed points: ones we never expect to withdraw, as when Lincoln says: 'If slavery is not wrong, nothing is wrong.' (…) The positions of judges, umpires, and referees are designed to include conditions that encourage the exercise of the judicial virtues, among them impartiality

> and judiciousness, so that their verdicts can be seen as approximating considered judgments, so far as the case allows. (Rawls, 2001, 29f.)

The conditions for considered judgments change throughout the development of Rawls' works. In his dissertation from 1950 and the essay *Outline of a Decision Procedure for Ethics*, he presents much stricter conditions for considered judgements (or, in this context, also "rational judgements"), requiring additionally that the epistemic agent forming the judgements should be a competent moral judge. This is not to be understood as an elitist position: Rawls thinks that any adult of intellectual integrity who has had some basic education is a competent moral judge, regardless of social, cultural or political background (cf. Rawls, 1950, esp. pp. 32–36; 1951). This condition for considered judgements is at least not mentioned explicitly any more from A Theory of Justice onwards. From Political Liberalism onwards, moreover, the focus moves away from considered judgements towards political judgements that are shared by reasonable comprehensive doctrines in an overlapping consensus. Especially in his later works, and in the political context of reasonable pluralism, he sees it as fruitful to start from commonly shared commitments (Rawls, 1999, p. 508; 2005, pp. 8–11; 2001, 31f.). However, the corresponding condition to include only those commitments in public justification which are shared in an overlapping consensus must be seen as distinct from the CJ-conditions.[3] For the purpose of discussing the plausibility of RE in general, we adopt the CJ-conditions as defined in *A Theory of Justice* and *Justice as Fairness,* as is commonly done.

One important aspect of the CJ-conditions is that they are intended to exclude obvious sources of error (cf. Daniels, 1979, p. 265), but Rawls does not claim that they are truth-conducive. Consequently, the CJ-conditions are a weak, not a strong, epistemic filter in the following sense (cf. Knight, 2017):

> *Strong epistemic filter*: We should only select commitments for which we can provide positive epistemic reasons.
> *Weak epistemic filter*: We should only select commitments that are free of obvious errors.

This means that commitments do not already possess some degree of justification through meeting the CJ-conditions—they have to be justified through being brought into a state of (wide) RE. Some critics of RE see this as a serious problem for RE, arguing that the process of adjustments cannot provide the necessary justification. Instead,

[3] For example, a utilitarian might have the considered judgement that an act is good insofar it maximizes the aggregated well-being. However, this judgement, while meeting the CJ-conditions for the utilitarian, arguably fails to be shared in an overlapping consensus.

they claim that a strong epistemic filter for commitments instead of the CJ-conditions would be needed, so that only commitments are admitted that already have some level of independent justification. We turn to this criticism in the next section.

2. Does Justification via RE Depend upon a Strong Filtering Process?

When the CJ-conditions are rejected as too weak, this is typically combined with a general criticism of reflective equilibrium being insufficient as a method of justification. According to these critics, even a correct and impeccable application of the method could lead to unreasonable and inadequate results. Reflective equilibrium could lead us to an epistemic position that is arbitrarily coherent without having any connection to the subject matter that it is supposed to represent (Brandt, 1979, p. 22; Kelly & McGrath, 2010, p. 333; Stich, 1993, pp. 83–86).

In other words, the criticism is that RE is too weak as a method of justification because it has no reliable mechanism to correct unreasonable input commitments—neither through the conditions for considered judgments, nor through the process of adjustments. This argument has been developed in detail by Kelly and McGrath (2010). They argue that something could qualify as a considered judgment without having any rational credibility. That this point is often missed is due to people using examples of considered judgments, which have some positive epistemic standing, even though this epistemic standing is not guaranteed through the CJ-conditions. As a commitment which meets the CJ-conditions while not having any rational credibility, they propose the example of the judgment "One is morally required to occasionally kill randomly" (Kelly & McGrath, 2010, p. 347). And they claim that not only could this bizarre judgment qualify as a considered judgement, it could even survive the process of adjustments and be part of a coherent position that qualifies as an RE state.

This is contrasted with applying our best scientific methods: If someone were to apply the best scientific methods to study fruit flies and started with various baseless assumptions about their nature, the correct application of the methods would guarantee that these baseless assumptions would be corrected or abandoned (Kelly & McGrath, 2010, pp. 327–328). Not so in the case of the RE, which—in the interpretation of Kelly and McGrath—aims only to achieve a coherent belief system by adjusting one's considered judgements. The idea seems to be that one could achieve an RE state by holding on to the *kill at random* judgement and adjusting all other commitments that are in conflict with it.

Kelly and McGrath conclude that, since the process of adjustments cannot guarantee that problematic commitments are excluded, one has to change the selection conditions for the input commitments. They therefore suggest abandoning the idea of considered judgements in favor of taking as appropriate starting points only

commitments that already have a certain positive epistemic status (Kelly & McGrath, 2010, p. 348). This is in line with Brandt's claim that RE can only lead to justified beliefs if at least some of the input commitments have an initial credibility that is independent of whether they are part of a coherent position (Brandt, 1979, p. 20).

According to Kelly and McGrath, however, this raises a serious problem for RE as a method of justification. If some input commitments were more rational or more justified than others, then this would have to be shown in some other way than through RE—thus, there would have to be some other kind of justification than the one RE can provide. That is, instead of a weak epistemic filter like the CJ-conditions, which only tries to exclude obvious sources of error, a strong epistemic filter for input commitments would be needed, which only admits commitments that have some positive epistemic status. From this they conclude that RE is not enough as a method of justification. The really interesting part is not the search for equilibrium, but the question which characteristics make certain input commitments more reasonable than others, and how we can recognize these characteristics (Kelly & McGrath, 2010, pp. 353–354).

As devastating as this criticism sounds, we argue that it rests on two misconceptions of RE: Firstly, the RE process remains underdeveloped, ignoring the holistic and social nature of RE, and secondly, the CJ-conditions are taken to be an essential element of RE. In the remainder of this section, we focus on the first misconception by bringing forward arguments against the need of a strong epistemic filter to fix RE. In sections 3 and 4 we will then discuss the status of the CJ-conditions as an element of RE, rejecting both that they are an essential element of RE as well as that they should be used for normative ethics.

We have two interconnected rejoinders against Kelly and McGraths argument that RE is implausible as a method of justification because a commitment such as the *killing at random* judgment could survive the RE process. First, the mere possibility that such a judgment could survive the process does not discredit RE as a method, and second, the RE process is much more demanding than what Kelly and McGrath seem to describe, making it extremely unlikely that something like the *killing at random* judgment could, in fact, survive.

First, why do we think that the mere possibility of such a judgment – from our perspective an obviously unreasonable one – surviving an RE process does not discredit RE? It is important to keep in mind that just because a principle or a judgement seems bizarre or absurd to us, it does not automatically follow that they are generally untenable or cannot be justified. Think, for example, of certain unintuitive principles of formal logic or principles of modern physics, such as Heisenberg's uncertainty principle (cf. Elgin, 1996, p. 119). We cannot know beforehand which judgments or principles would result as justified, and excluding initially implausible judgments could lead to conservatism and hinder progress by making RE less revisionary (cf. Dutilh Novaes, 2020, Fn. 5). However, surviving the RE process and being part of a wide RE state is more demanding than what Kelly and McGrath's description

suggests: If a commitment such as the *kill at random* judgment—against all expectations—were part of a wide RE, we would be able to explain why our initial assessment of this judgment as absurd was wrong, and how it can be defended. RE in this sense has a strong explanatory dimension.[4] Then, although the commitment seemed quite bizarre and unjustified at the beginning, it would actually be reasonable to hold for us. This leads us to our second, more extensive, rejoinder.

Second, it is highly unlikely that a commitment such as the *kill at random* judgment would survive a real process of adjustments that aims at wide RE. What is striking about their argument is that Kelly and McGrath focus so strongly on the starting conditions that the very idea of RE hardly matters anymore. They seem to think that the resulting set of principles, or the resulting theory, will be pretty much a direct expression of the input commitments, i.e., the considered judgments of the epistemic agent. However, this is very unlikely, especially in the case of wide RE (see also Knight, 2017, p. 53). As mentioned in section 2, a wide RE requires that all relevant philosophical arguments have been considered and that the resulting epistemic position is more plausible than all relevant alternative positions. This includes that not only should one's own commitments be made coherent with each other, but that, for example, the principles chosen should additionally be supported by background theories which themselves ideally should have some additional support (Daniels, 1979, pp. 259–260). Importantly, adjusting one's commitments in light of all relevant considerations does not just mean "considering whether I want to accept that", and being allowed to stick to specific commitments no matter what. For justification we cannot simply decide to disregard the real strength of our own commitments and the inferential relations between them, including conceptual connections, or reasons speaking in favor or against a specific commitment. Applied to the *kill at random* judgment: Even if there were a person who would be committed to this judgement, there are good reasons that speak against it, and those would have to be duly considered as part of the process, and cannot simply be disregarded without being refuted.

Now, the critics might answer that of course I can stick with a specific commitment like *kill at random* no matter what: The only real criterion of RE seems to be coherence, and I could refute reasons against a commitment that I want to keep based on the argument that my epistemic position is more coherent without accepting those reasons as valid. But this paints again an oversimplified picture of RE and its requirements: Not just *any* coherent set of commitments will do. As Rawls emphasizes, a list of our commitments, even if they were consistent with each other, is not enough. To really have an adequate account of a subject like justice, we need principles that *systematically* cover a wide range of cases and show how the different judgements relate to each other (e.g., Rawls, 1999, p. 41). Moreover, one has also to consider theories, alternative principles and commitments which are being put forward by

[4] We thank an anonymous referee for stressing this point.

epistemic agents one judges to be epistemic peers or experts with regard to the subject matter.[5] The corresponding drive towards systematization is one of the forces in RE that counteract conservatism with respect to our input commitments.[6] The search for systematic principles that are supported by background theories and that can account best for our commitments favors abandoning idiosyncratic judgements for which we cannot find other convincing reasons. Conversely, of course, it speaks against a principle if it conflicts with many of our plausible commitments and cannot be additionally supported by other considerations.

Additionally, coherence is more than consistency, and both coherence and systematicity are a matter of degree. Consequently, we can compare possible resulting RE states and evaluate them with respect to their degree of coherence between commitments and principles as well as the degree to which they are supported (or undermined) by background theories. Moreover, and importantly, coherence is instrumentally important for the requirement that our epistemic position should be *the most plausible or at least as plausible as other relevant available alternatives*. This means that in cases of conflict, we have to consider what speaks in favour or against certain (sets of) commitments, and make adjustments in a way that makes the position as a whole more credible. As coherence involves interconnectedness and mutual support, sticking with idiosyncratic commitments and making ad hoc adjustments to save them will typically cause problems for the overall credibility of our epistemic position.

Consequently, even though Kelly and McGrath distinguish two ways how an input commitment might be adjusted, they concentrate on the filtering through the CJ-conditions and do not sufficiently pay attention to the requirements of wide RE.[7] How a real RE process would look like which starts and ends with the *kill at random* commitment, remains as open as what untenable assumptions the person in the fruit fly example has and how these are filtered out by applying the "best scientific methods" (cf. Knight, 2017, pp. 53–54). We can, of course, not prove that it is impossible for such a commitment to survive a process of adjustments and end up as part of a position that is in wide RE. However, we hope to have made plausible that it is highly unlikely that a commitment like "One is morally required to occasionally kill randomly" would be part of a wide RE, i.e., an epistemic position that is at least as plausible as relevant alternatives that do not include this commitment.

---

[5] Considering alternative principles of justice that are debated in moral philosophy is one key element of an adequately wide RE concerning justice according to Rawls (1999, p. 43). And Elgin shows that an adequately wide RE includes collaboratively gained rules, inferences and grounds that would lead to identify the Gambler's Fallacy as a fallacy, contra Stephen Stich (Elgin ,1996, 118f.)

[6] For further elaborations on the role of systematisation for justification via RE, see Brun (2020).

[7] Additionally, it is not clear whether the *kill at random* commitment would really meet the CJ-conditions, e.g., weak impartiality—would the epistemic agent also accept that they could be killed randomly at any time? However, for the sake of discussing the resulting argument, we grant this point to Kelly and McGrath.

This can be further illustrated with an example from Daniels, who discusses the judgment "It is wrong to inflict pain gratuitously on another person". Like everything that enters an RE process, this commitment is not principally save from revision. But as Daniel explains, RE allows us to explain why it is unlikely that we would completely abandon it:

> Since all considered judgments are revisable, the judgment "It is wrong to inflict pain gratuitously on another person" is, too. But we can also explain why it is so hard to imagine not accepting it, so hard that some treat it as a necessary moral truth. To imagine revising such a provisional fixed point we must imagine a vastly altered wide reflective equilibrium that nevertheless is much more acceptable than our own. For example, we might have to imagine persons quite unlike the persons we know. (Daniels, 1979, p. 267)

Analogously, reflectively accepting the *kill at random* commitment would require drastic changes in our system of commitments: Imagine, for the sake of argument, that the commitment itself has an initial acceptance, although it normally invokes an immediate virtually ultimate rejection. However, there are many arguments ready at hand that can support its rejection; arguments based on premises that we are firmly committed to and that are themselves well-supported by many case-based judgements, like the validity of human rights or some version of the *Golden Rule* (see Parfit, 2011, pp. 321–330). Yet, it is very hard to imagine at least one argument that supports the commitment and has premises that all have initial credibility themselves. Thus, even in the counterfactual situation that we accept the *kill at random* commitment, it would be certainly outweighed – no matter how strong we imagine the initial acceptance of it. This evaluation is, of course, based on empirical assumptions about the moral commitments people actually have. However, even though philosophers should be extremely cautious about making empirical judgements from the armchair position, this judgement—even considering the cultural plurality our world inhibits—seems not to be a very bold one.

These rejoinders counter the criticism that RE cannot correct unreasonable input commitments. It is, of course, always possible for someone to insist on a commitment and refuse to see that their position would be more plausible without it. But contrary to Kelly and McGrath's assertion, the method of the wide RE does provide us with the means to criticize such a person's epistemic position.[8]

Thus, we reject the claim that RE would need a strong epistemic filter in the first place, which only admits commitments with some positive epistemic standing into the RE process. We thereby answered one line of criticism

[8] As our first rejoinder showed, we concede the following: If such a judgement nevertheless would turn out to be a part of an *adequately wide RE state* for an epistemic agent, this simply would mean that it is—after all—reasonable to accept for this specific agent. Yet, this would not hinder other epistemic agents to obtain a different RE for themselves and to act accordingly.

the CJ-conditions face, namely, that they are too weak, i.e., too inclusive. However, they are also being criticized as too strong, i.e., too exclusive (in fact, also by Kelly & McGrath, 2010, p. 348). We will address this objection in two steps: First, we will address the second misunderstanding on which the criticism of Kelly and McGrath rests, namely, we argue that the CJ-conditions are not an essential part of RE. This means that even if they are too exclusive, this does not cause a problem for RE as a method of justification in general. Then, in section 4, we will turn to the question whether using the CJ-conditions as a weak epistemic filter can be justified for the context of normative ethics—arguing that they are, in fact, too exclusive and should be abandoned.

3. Is the Weak Rawlsian Filtering Process an Essential Element of the General RE Method?

Critics like Kelly and McGrath reject RE as a general method by focusing on the CJ-conditions, taking them to be an essential element of RE. Those conditions, however, originate from the Rawlsian conception of RE for moral and political philosophy. This section develops an argument that RE as a general method does not depend on specific conditions for pre-selecting commitments—and we argue that even Rawls could agree with this. However, as we will see, it remains possible that such filter-conditions for commitments can be justified for specific projects and contexts.

Assuming that the CJ-conditions are a definitely fixed element of RE as a general method of justification misses the holistic and dynamic character of RE. Every element of an epistemic position that is in a state of wide RE is ultimately justified through its connections to all other elements of the position. Importantly, this also holds for the criteria and constraints placed on the elements of the position, like theoretical virtues of principles, or conditions for considered judgments (cf. Scanlon ,2016, pp. 83–86; Elgin, 1996, p. 105; 2017, pp. 89–90). Thus, if it turned out that Rawls' conditions for considered judgments are unsuitable—for example, because we arrive at less plausible systems of commitments than without them—this would not be a reason to reject RE as a method in general.

We therefore propose that even proponents of RE that want to use a *filter* like the CJ-conditions for initial commitments should not regard these criteria for pre-selecting initial commitments as a definitely fixed element of RE. Their position becomes more plausible if they concede that filtering criteria should themselves be justified by an application of the RE-method—and that these criteria are, like everything in RE, open to revisions and adjustments.

Let us elaborate. Even though justification via the RE method is fundamentally holistic, it often makes sense to take certain aspects in the background as given while we focus on a particular justification project in the foreground. This is especially the case if we can show these background assumptions, theories, or criteria to be justified, e.g., through their own RE-process. That is, filter criteria for pre-selecting initial commitments can be used if it can be

shown that they are part of their own wide RE state—and if they are not regarded as fixed once and for all, but as subject to critical reflection and possible revision. Importantly, an attempt at justifying a set of filtering conditions via RE does not itself presuppose criteria for filtering initial commitments. Instead, specific criteria (like a filter) could be justified for specific areas of inquiry via a more general application of RE.

Consequently, criteria for pre-selecting commitments are not a necessary part of the general method of RE. This is plausible because the main RE process of mutually adjusting commitments and systematic principles remains functional without a filtering process. Not seeing the CJ-conditions or other criteria for pre-selecting initial commitments as essential to RE thus avoids circularity and results in a higher critical potential of the method.

Notably, we argue that according to our interpretation, even Rawls and Daniels could agree with this conclusion. Both see the method of RE as a general method of justification (Rawls, 1999, 17ff.; Daniels, 2020; but cf. Freeman, 2007 for an opposing view) This makes it unproblematic for Rawls and Daniels to accept the conclusion, since they do not have to understand the criteria for pre-selection of initial commitments as an essential element of the method, but only do prescribe it as binding with respect to the special area of normative ethics and, more specifically, political philosophy.

This is in line with our argument that specific criteria can be justified for special areas of inquiry via a more general application of RE. For example, with regard to empirical investigations it might be possible to justify a moderate foundationalism according to which perceptual beliefs are to be regarded as basic beliefs that are contestable but do not require further justification by other beliefs. The further differentiation of special methods of investigation and confirmation in different scientific disciplines could also be justified in this way by applying RE. Like everything justified by RE, these criteria and methods are of course not ultimately justified, but can be "unbalanced" at any time by new findings or insights. Justification via the method of RE is plausible and reasonable, but fallible (see also Elgin ,1996, p. 15, pp. 133–34; 2017, pp. 89–90)

To sum up, even if the CJ-conditions turned out to be untenable, this would not undermine RE as a method in general: First, the interpretation of RE and its elements is always open to revision, nothing is definitely fixed. Second, the CJ-conditions should not be interpreted as an essential element of RE anyhow (cf. Walden, 2013 for the stronger position that we should resist essentialist interpretations of specific RE criteria in general). However, the question remains whether they can be defended for the context that Rawls and Daniels intended them for, that is, normative ethics and political philosophy. We turn to this question in the next section.

4. Can a weak Rawlsian epistemic filter be justified for normative ethics?

We have argued that a) the CJ-conditions are not an essential element of the general method of RE, and that b) to be a justified element of a specific conception of RE, they would need to be justified themselves. In this section, we ask: Can the CJ-conditions (or another weak epistemic filter) be justified for normative ethics, i.e., the context for which Rawls developed his conception of RE? Justifying a weak filter like the CJ-conditions would require to show that these criteria for the pre-selection of input commitments are part of a position that is in a state of wide RE with respect to the justification of commitments in normative ethics. In the following, we survey some reasons for and against the CJ-conditions in this context. While we can of course not present a final verdict, we argue that there are good reasons to abandon even a weak epistemic filter such as the CJ-conditions.

Let us start by considering what speaks in favor of using the CJ-conditions to preselect input commitments. As a first point, we can note that to exclude obvious sources of error has some prima facie plausibility. The CJ-conditions are common factors to which we refer when explaining why someone has come to hold incorrect beliefs (Rawls, 1999, p. 42). Thus, if a commitment is based on negative factors like self-interest, it can be seen as decreasing the coherence of our system as such factors detract from the credibility of the individual commitment—and thereby indirectly from the credibility of the position that the commitment is a part of. Excluding them from the start thus promises to increase the overall credibility of our epistemic position (Tersman, 1993, p. 49).

Additionally, using the CJ-conditions (or a similar weak epistemic filter) can be defended as fulfilling a heuristic function in situations where we have to conduct inquiry under time constraints. If we only take into account commitments that seem free from obvious sources of error, we have less commitments to vet during the process. Arguably, this simplification is likely to allow us to reach an equilibrium faster, and makes it less likely to have biased commitments distort the process, leading us on tangents, and running out of time before we reach even a preliminary, defensible solution.

Lastly, we can also make a more substantial, moral argument at least for the condition that the epistemic agent should be impartial (in a weak sense) and thus not be influenced by personal interests. This condition can be seen as supported by certain expectations about what makes a judgment a *moral* judgment: One could argue that moral judgements are inherently characterized by such impartiality (cf. Jollimore, 2021). Judgments that were not made from this impartial perspective would thus simply not be moral judgments and could accordingly be considered irrelevant for investigations in moral philosophy. Even if the filter was justified in this way and thus would imply certain assumptions about morality, this would not amount to *question-begging*: The weak impartiality condition would not yet pre-decide that moral judgements are impartial *only* in the sense of the original position—under the veil of ignorance—and therefore necessarily in accordance with Rawls' principles of justice. Utilitarianism, for

example, also advocates forms of impartiality. Rawls explicitly mentions the impartial sympathetic spectator in this context, as a serious alternative conception to the original position (Rawls, 1999, pp. 161–163). Accordingly, intuitions that support a utilitarian moral theory, or a utilitarian theory of justice, can also meet the conditions for considered judgments.

However, we argue that these reasons are not enough to establish the considered judgment conditions—or any other form of weak epistemic filter—for normative ethics. Even the seemingly minimal conditions for considered judgments can still be too exclusive, and this also holds for other forms of a weak epistemic filter that would aim to exclude erroneous commitments before even entering the process of adjustments in RE. In the following, we look at four arguments against a weak epistemic filter for normative ethics. The first three address specifically the considered judgments conditions, while the fourth is a general rejection of pre-selecting moral commitments according to their epistemic standing in any form.

Firstly, the impartiality condition can be interpreted in a way that excludes members of marginalized groups from taking into account their own experiences of injustice and discrimination when applying RE. Kelly and McGrath discuss the example of the proposition 'A person of color should not receive lesser consideration in virtue of being a person of color.' They conclude that because this judgment will be heavily bound up with the personal interests of a person of color, it seems like it would not qualify as a considered judgment for this person—and consequently would be excluded from her application of RE. But this seems like the wrong result (Kelly & McGrath, 2010, p. 348). While this particular example seems to rest on an uncharitable interpretation of Rawls, it still points to a relevant problem. It is not clear at all that the condition of weak impartiality would exclude such and similar judgments based in personal experiences of injustice. It seems more plausible to interpret the condition as requiring that the person in question regards the judgment as neutrally as possible, and that they ask themselves whether they would also agree to the judgement if they were not personally affected—or if they are endorsing it *exclusively* because of their own interests. According to this interpretation, the judgment would only in the latter case not qualify as a considered judgement.

Nonetheless, the question remains whether each commitment should first undergo such a reflection before being admitted as input to the process of adjustments. Should not even the immediate and perhaps very existential judgment "I am being treated unfairly here" be considered relevant in moral and political reflection and theorizing? This is especially true in view of the fact that the judgment, should it ultimately turn out to be unjustified, can only be modified or revised through the RE process if it continues to be considered in the process. Whether or not such a judgment rests purely on self-interest, or whether it can be vindicated as part of a wide RE state, is not something that

can be decided prior to inquiry. If we exclude such commitments, we might exclude relevant information from the process.

Secondly, the same line of reasoning—that a preselection could exclude relevant considerations from the process—also applies to the condition that only judgments made in a calm mood should be considered. Recent work in the philosophy of emotions argues that emotions can play the role of defeasible reasons for normative judgments (one example is Tappolet, 2016). Emotions can also play an epistemic role by pointing us to certain aspects that we overlook or repress when we are in a calm and composed state. In this way, emotions can make us aware of injustices that we otherwise would overlook because we are lacking the conceptual tools for naming and describing them, that is, due to what Fricker (2007) calls "hermeneutical injustice" (see Scheman, 1980 for an instructive example how emotions can point to an injustice which cannot be articulated at the time). Consequently, excluding commitments previous to the process of adjustments just because they are based on strong emotions does not seem warranted (for the role of emotions in RE, see DePaul, 1993, pp. 180 et sqq.; Elgin, 1996, Chapter V).

Thirdly, the same objection applies to the condition that only judgments that we hold confidently should be considered. If we want to make moral progress, then at least in some cases the judgments that we only make reluctantly or with little confidence will be the most progressive ones. Only because we have little confidence in them at the beginning does not mean that they cannot be vindicated through being brought into a position that is in reflective equilibrium. This point is nicely illustrated by the following quote from de Maagt:

> […] it might well be that in times when homosexuality was commonly regarded as a sin, someone who judged the opposite was in fact hesitant about this judgment. Surely, we do not want to exclude these kind of progressive moral judgments from our moral methodology simply because people were upset or uncertain about their judgments. (de Maagt, 2017, p. 453)

This leads us to the fourth, more general point, that a preselection of commitments according to the considered judgments conditions or another weak epistemic filter could undermine the dynamic and critical character of RE. For example, if an epistemic agent already preselected their commitments as to only include those that they are relatively confident in, they might be more reluctant to make adjustments. Additionally, as stressed by van Thiel and van Delden (2009, p. 235), applying RE is also in part a creative process that relies on having a broad range of input which can fuel the thinking process. Agents should thus strive to continuously broaden the set of commitments they consider, not make a preselection in the beginning.

To summarize, it contributes to the innovative power of RE to allow input commitments that only have a very minimal independent credibility—i.e., which might be nothing more than a hunch. Many of these low credibility commitments might be rejected relatively quickly, and considerations such as whether they are impartial, have been made in a calm mood, or how confident we are in them, etc., will and should play a role when deciding which commitments to adjust and which to keep. However, it can also be the case that a combination of different commitments with low independent credibility taken together outweigh commitments with a high independent credibility (see, e.g., Elgin, 2017, p. 69). Not pre-selecting commitments through epistemic filter conditions is thus one aspect that helps us to use RE to correct prejudices and to go beyond what we are already convinced of.

It does not follow that the reasoning behind the CJ-conditions becomes irrelevant. In fact, whether or not a commitment is based on self-interest, or strong emotion, etc. are relevant factors—which should be considered as part of the RE process, and not be implemented in form of a previous selection.

Nonetheless, justification via RE is not justification ex nihilo: As the remaining minimal requirement to be considered in the process of adjustments, the epistemic agent needs to have at least a minimal degree of *commitment* to the judgment in question and is willing to accept it at least as a working hypothesis. Completely hypothetical, made-up, implausible *judgments* that nobody is committed to can and should still legitimately be excluded. The question which of these commitments should be rejected or revised as erroneous, and what the criteria are for making these decisions, is, however, a result and not a precondition of applying RE (cf. Baumberger & Brun, 2021).

5. Conclusion

This paper discussed the role of the Rawlsian CJ-conditions for the method of reflective equilibrium. We have explored, firstly, whether RE needs a stronger epistemic filter than the CJ-conditions, secondly, whether the CJ-conditions—or a similar weak epistemic filter—should be seen as an essential part of RE, and lastly, whether the Rawlsian CJ-conditions can be justified for inquiries in moral philosophy. We answered all three in the negative.

We have argued that justification via RE does not depend on an (external) strong epistemic filter. Indeed, if epistemic agents do apply the method successfully, the holistic perspective and rational requirements of RE have enough force to discard unreasonable commitments and provide an internal justification for the accepted ones (although it does not guarantee the truth or correctness of the commitments).

Although a filter—strong or weak—might indeed be justified for specific areas of inquiry, their justification would depend on being included in a more general RE that does not presuppose the filter. Consequently, filtering processes for input commitments are not an essential element of the method of RE.

Concerning the question whether a weak Rawlsian filter in the form of the CJ-conditions is justified for inquiries in moral philosophy, we are skeptical. As we have shown, there are serious objections to them, and they might be too exclusive. Since all the effect of the filter can be gained by simply including the arguments that support it in the RE process without excluding commitments that do not fulfill the CJ-conditions, rejecting the filter seems to come with less epistemic risk and thus is more adequate. We leave it open whether the CJ-conditions might be justified for inquiries with serious time constraints as a heuristic function, but would hold that this is typically not the case in moral philosophy.

Although we criticize an element of the Rawlsian conception of reflective equilibrium, we would like to conclude by highlighting that this critique does not necessarily abandon the Rawlsian spirit of the method, but is compatible with the rest of his theoretical framework. We suggest that the revisions corresponding to our criticism can be interpreted to be in line with an aspect of the historical development of reflective equilibrium in Rawls' oeuvre. His focus shifted from even stronger CJ-conditions, that included a filter for commitments originating from competent moral judges (Rawls, 1950; 1951), to the distinct requirement that justification in the political realm must rest on commitments which are shared in an overlapping consensus (Rawls, 2005; 1995; 2001). Abandoning the element of filtering input commitments through the CJ-conditions could thus be seen as a radical continuation of Rawls' own tendency to alter the conditions for commitments to enter the RE process. If one wanted to keep the term *considered judgment*, then we suggest to use it for those judgments that result from a successful RE process—that is, that have been duly considered in light of all relevant considerations. And the latter can, of course, include considerations from the original CJ-conditions—just not in the form of a preliminary filtering.

Acknowledgements

Previous versions of this paper have been presented online in Braga and Karlsruhe. We thank the respective audiences and discussants for their helpful and constructive feedback. Additionally, we would like to thank Thorben Knobloch, Richard Lohse, Irina Schumski, and an anonymous reviewer for their critical questions and insightful comments.

References

Baumberger, C., & Brun, G. (2021). Reflective equilibrium and understanding. *Synthese*. 198(8), 7923–7947. https://doi.org/10.1007/s11229-020-02556-9.

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

Brun, G. (2013). Reflective equilibrium without intuitions?. *Ethical Theory and Moral Practice*, 17(2), 237–252. https://doi.org/10/gf7fn4.

Brun, G. (2020). Conceptual re-engineering: From explication to reflective equilibrium. *Synthese*, 197(3), 925–954. https://doi.org/10.1007/s11229-017-1596-4.

Daniels, N. (1979). Wide reflective equilibrium and theory acceptance in ethics. *The Journal of Philosophy*, 76, 256–282.

Daniels, N. (2020). Reflective equilibrium. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy (Summer 2020 Edition)*. https://plato.stanford.edu/archives/sum2020/entries/reflective-equilibrium/.

DePaul, M. R. (1993). *Balance and refinement beyond coherence methods of moral inquiry*. Routledge.

de Maagt, S. (2017). Reflective equilibrium and moral objectivity. *Inquiry*, 60(5), 443–465. https://doi.org/10/ghmqck.

Dutilh Novaes, C. (2020). Carnapian explication and ameliorative analysis: A systematic comparison. *Synthese*, 197(3), 1011–1034. https://doi.org/10/gmpp6s.

Elgin, C. Z. (1996). *Considered judgment*. Princeton University Press.

Elgin, C. Z. (2017). *True enough*. The MIT Press.

Freeman, S. R. (2007). *Rawls*. Routledge.

Fricker, M. (2007). *Epistemic injustice: Power and the ethics of knowing*. Oxford University Press.

Jollimore, T. (2021). Impartiality. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy (Fall 2021)*. https://plato.stanford.edu/archives/fall2021/entries/impartiality/.

Kelly, T., & McGrath, S. (2010). Is reflective equilibrium enough?. *Philosophical Perspectives*, 24(1), 325–359. https://doi.org/10/bmh3g5.

Knight, C. (2017). Reflective equilibrium. In A. Blau (Ed.), *Methods in Analytical Political Theory* (pp. 46–64). Cambridge University Press.

Parfit, D. (2011). *On what matters*. The Berkeley Tanner Lectures. Oxford University Press.

Rawls, J. (1950). *A study in the grounds of moral knowledge: Considered with reference to the moral worth of character*. Princeton University.

Rawls, J. (1951). Outline of a decision procedure for ethics. *Philosophical Review*, 60(2), 177–197. https://doi.org/10/fpn9w7.

Rawls, J. (1974). The independence of moral theory. *Proceedings and Addresses of the American Philosophical Association*, 48, 5–22. https://doi.org/10/b6skdk.

Rawls, J. (1995). Political liberalism: Reply to Habermas. *The Journal of Philosophy*, 92(3), 132. https://doi.org/10.2307/2940843.

Rawls, J. (1999). *A theory of justice* (Revised ed.). Belknap Press.

Rawls, J. (2001). *Justice as fairness: A restatement*. Belknap Press.

Rawls, J. (2005). *Political liberalism* (Expanded ed.). Columbia University Press.

Rechnitzer, T. 2022. *Applying Reflective Equilibrium. Towards the Justification of a Precautionary Principle. Logic, Argumentation & Reasoning*. Cham: Springer. https://link.springer.com/book/9783031043321.

Scanlon, T. (2016). *Being realistic about reasons*. Oxford University Press.

Scheman, N. (1980). Anger and the politics of naming. In N. Furman, R. Borker, and S. McConnell-Ginet (Eds.), *Women & Language in Literature & Society* (pp. 22–35). Praeger.

Schmidt, M. W. 2022. *Das Überlegungsgleichgewicht als Lebensform: Versuch zu einem vertieften Verständnis der durch John Rawls bekannt gewordenen Rechtfertigungsmethode*. Brill mentis. https://brill.com/view/title/61282.

Stich, S. (1993). *The fragmentation of reason: Preface to a pragmatic theory of cognitive evaluation*. MIT Press.

Tappolet, C. (2016). *Emotions, value, and agency*. Oxford University Press.

Tersman, F. (1993). *Reflective equilibrium: An essay in moral epistemology*. Lagerblads tryckeri AB.

van Thiel, G. J., & van Delden, J. J. (2009). The justificatory power of moral experience. *Journal of Medical Ethics*, 35(4), 234–237. https://doi.org/10/fjpdbg.

Walden, K. (2013). In defense of reflective equilibrium. *Philosophical Studies*, 166(2), 243–256. https://doi.org/10/gf7fn5.