Reflective Equilibrium*

Carl Knight, University of Glasgow

1. Introduction

The method of reflective equilibrium focuses on the relationship between *principles* and *judgments*. Principles are relatively general rules for comprehending the area of enquiry. Judgments are our intuitions or commitments, 'at all levels of generality' (Rawls 1975: 8), regarding the subject matter. The basic idea of reflective equilibrium is to bring principles and judgments into accord. This can be achieved by revising the principles and/or the judgments. For instance, if I am considering the principle that it is always wrong to lie, but have the judgment that it would not be wrong to lie in order to save a life, I can reach equilibrium by either revising the principle or revising the judgment.

Reflective equilibrium is the most widely used methodology in contemporary moral and political philosophy (Sinnot-Armstrong et al 2010: 246; Varner 2012: 11). It has even been suggested that it is 'the only defensible method' (Scanlon 2003: 149). Its popularity is undoubtedly strongly influenced by John Rawls' use of it in his seminal *A Theory of Justice,* published in 1971.[1] However, the method precedes this, and extends to other fields. For instance, Nelson Goodman wrote regarding induction that '[t]he process of justification is the delicate one of making mutual adjustments between rules and accepted inferences; and in the agreement reached lies the only justification needed for either' (Goodman 1965: 64). Some fields, by contrast, do not seem as amenable to the method of reflective equilibrium. Within linguistics, for instance, native speakers' judgments of grammaticality cannot generally be replaced as moral judgments can (Daniels 1996: ch. 4). Although most writers treat reflective equilibrium as unproblematic within empirical sciences (Daniels 1996: 31-33; Cummins 1998; Welch 2014: 4; see also McDermott 2008), adjustment of empirical judgments also seems to be subject to stronger constraints than those that apply to moral judgments (see Singer 2005: 345).

Indeed, here there seems to be a significant difference between normative political theory and empirical political science. A normative political theorist who, to her surprise, finds that a confidently held moral judgment conflicts with an otherwise compelling principle (or set of principles) is free to reject that judgment precisely because it conflicts with the favoured principle. But it would be quite improper for a confidently-held empirical judgment to be abandoned simply because it turned

* In Adrian Blau (ed.), *Methods in Analytical Political Theory* (Cambridge University Press, 2017).

[1] Jo Wolff (2013: 808) notes that, of the papers collected in the first two series of *Politics, Philosophy, and Society,* Rawls' was unique in aiming to defend a substantive position, and in deploying a distinctive methodology to positive effect.

out not to fit with the investigator's pet hypothesis. Full reflective equilibrium, with judgments adjusted at will just as principles are, is primarily the reserve of normative political theory. It is with normative political theory that this chapter will be concerned.

I first look at normative political judgments (section 2) before considering the role of principles, arguments, devices of representation and background theory in wide reflective equilibrium (section 3). I then consider two of the main challenges to the method (section 4), and show how to use it to deliberate about substantive political principles (section 5). I conclude with an extended example of the method in action (section 6).

2. Judgments

The starting point for reflective equilibrium is our judgments. We cannot, however, use just any judgments. For instance, judgments made 'in the heat of the moment' would not be a reliable basis for equilibrium. *Considered* judgments are what we need. These are 'those judgments in which our moral capacities are most likely to be displayed without distortion' (Rawls 1999: 42).

Most writers, including Rawls (1999: 42), suppose that, in order to count as considered, judgments should be held with *confidence*. Indeed, Rawls often uses 'convictions' as a synonym for 'judgments' (Rawls 1975: 8; 2005: 24, 26, 28, 151, 156). This 'confidence constraint' seems to me quite gratuitous (Knight 2006: 207-208). If I have the firm conviction that the state should protect its citizens from terrorism, and also believe, less firmly, that individuals have a right to privacy, a right to not be subject to pre-trial detention beyond a certain duration, and a right to a public trial, the confidence constraint would seem to require that my numerous but less firmly held concerns about individual rights be set aside. But this is to give free reign to the one firm conviction, with the upshot that the principle(s) arrived at in reflective equilibrium will allow almost any breach of civil rights in the name of public protection. This is in contradiction of the majority of the judgments I hold and (most likely) my overall view.

Undoubtedly, a firmly held judgment should generally carry more weight in our deliberations than a less firmly one. But reflective equilibrium *automatically* does that, as we are presumably more likely to give up our less firmly held judgments in the face of opposing judgments or principles. If we don't, that's because it turns out that the less firmly held judgments had something going for them. Maybe they individually or collectively capture something that, on reflection, we consider to be important. Thus, I think we should reject the confidence constraint.

A different constraint has sometimes been assumed, including in my earlier work (Knight 2006: 207). This specifies that our considered judgments do not display

errors of reasoning, such as logical inconsistencies, or empirical errors. Some writers go further, suggesting that we should disregard judgments that we don't have evidence for, as they are not epistemic assets (Gaus 1996: 86; Kelly and McGrath 2010: 347-354). Should we, then, endorse an 'epistemic constraint', requiring that only justified or warranted judgments, or (more minimally) only those that lack errors, are admitted to the reflective equilibrium procedure?

Though this may seem like simple common sense, I doubt it. Consider first the stronger version of the constraint, which requires justification or warrant for a judgment. Evidence can be rather thin on the ground when we are dealing even with firmly held judgments. If I consider some political judgment that I hold very firmly, such as the judgment that no fellow citizen should avoidably starve, it is hard to point to anything that can really count as evidence for that belief in the relevant sense. I might point out that my compatriot will suffer pain, reduced capability, and eventually death, but these *empirical* facts alone cannot really be evidence for the *normative* judgment I am making. It seems that, in a case like this, the judgment itself is foundational. My judgment seems pre-theoretically plausible to me, and that is sufficient to grant it 'independent credibility' (Hooker 2012: 23). This does not mean that it has any weight in my final principles, but it is enough for it to be granted admission to the reflective equilibrium process. It is there that the credentials of our judgments are really tested, by seeing how well they fit with our other judgments and the most plausible principles, in light of the most compelling arguments we can muster.

What then of the less demanding version of the epistemic constraint, which requires merely the absence of outright error? Surely we can reject some judgments as clearly erroneous. But even this constraint might be thought to be excessively demanding in that it goes beyond providing 'conditions favorable for deliberation and judgment in general' (Rawls 1999: 42) and actually limits the admissible content of judgments. Furthermore, exactly what qualifies as an error of reasoning and what qualifies as an empirical error is controversial. We could consider these issues in piecemeal fashion prior to entering reflective equilibrium. But this is counterproductive as we have no way of knowing whether these isolated speculations will be consistent with the most plausible overall position. We should instead consider these issues holistically, as pieces in the jigsaw that is the coherent view of the conceptual terrain that we aim to arrive at in reflective equilibrium. (Specifically, these issues are settled through consideration of relevant background theories – see section 3 below.) Reflective equilibrium eschews the essentialist notion 'that we can determine the nature of certain facets of these inquiries in advance of the inquiries themselves, and that nothing that comes about in inquiry will change those facets' (Walden 2013: 255). The epistemic constraint, even in its minimal form, seems to put the cart before the horse, and should be discarded. Considered judgments are just those made in 'conditions favorable for deliberation and judgment in general'.

3. Wide reflective equilibrium

Suppose that you have arrived at your set of considered judgments. You might first use these to reach *narrow* reflective equilibrium, in which 'one is to be presented with only those descriptions which more or less match one's existing judgments except for minor discrepancies' (Rawls 1999: 43; see also Rawls 2005: 8 n. 8). Narrow reflective equilibrium is in essence an effort to systematize an agent's pre-theoretical views. As such, it has limited epistemic value. Were someone to ask you what justification you have for your principles, you do not have much of a reply. To be sure, the narrow reflective equilibrium principles might be an improvement from your perspective on the bare intuitions you started out with. But you can hardly say that your principles are well justified where they are just a direct expression of your pre-theoretical intuitions.

The more interesting version of the method is *wide* reflective equilibrium. Rawls describes this in very demanding terms: 'one is to be presented with all possible descriptions to which one might plausibly conform one's judgments together with all relevant philosophical arguments' (Rawls 1999: 43). So for wide reflective equilibrium to be reached, you must consider all principles (and combinations of principles) that you might accept. As one way in which you may conform your judgments to principles is to change your judgments, this means that you must consider every principle in every combination with every other principle!

Unsurprisingly, Rawls does not attempt to fully satisfy this unachievable standard, resolving in *A Theory of Justice* to compare only his own 'principles and arguments with a few other familiar views' (Rawls 1999: 43). For all practical purposes, it will undoubtedly be necessary to narrow our equilibrium in this way. Nevertheless, I think there is great value in keeping in mind that wide reflective equilibrium is an *ideal*. It sets the bar high. Though the theorist will inevitably only consider a few principles, this is not because that is all the method of reflective equilibrium requires for a full justification to be provided. It should always be kept in mind that consideration of more principles would provide a fuller justification. Furthermore, if we have to cut corners, we should do so in the way least harmful to the strength of the final justification. This means, for example, ensuring that we at least consider the most compelling rival principles, rather than satisfying ourselves with seeing off straw men.

Reflective equilibrium can be interpreted as providing an ecumenical answer to a long-standing problem in epistemology. The ordinary way of justifying beliefs is *inferential* and *linear:* belief A justifies belief B, which justifies belief C, and so on. The problem here is rather obvious. As the chain of inference cannot go on infinitely, it seems that *none* of our beliefs will be justified. There are two ways out of this infinite justificatory regress. *Foundationalism* denies that all justification is inferential – for example, A might be justified by something other than another

belief. *Coherentism* denies that all justification is linear – for example, C might be justified by A (Brink 1989: 109). A large majority of writers see reflective equilibrium as a coherentist method (Brink 1989: 134; Daniels 1996: 60-61; Tersman 2008: 398-400; Maffetone 2010: 142-145), while a few see it as foundationalist (DePaul 1986; Ebertz 1993). In my view it clearly contains elements of both approaches. Foundationalism can be seen to be present as I would, according to the method, be justified in favouring one possible coherent set of principles and considered judgments to another purely because the former coincides with my actual considered judgments. Coherentism cannot explain this, as each set is identical as regards coherentist non-linear justificatory chains. But coherentism is evidently also present, as the method says that the fact that some judgment (or principle) coheres with the rest of our beliefs counts in its favour.

As I have mentioned, Rawls requires not only that principles be considered, but relevant arguments as well. We could reach equilibrium without arguments, but coherence among beliefs that have not been subjected to serious scrutiny would be of limited justificatory value. This introduces several new complexities. First, there are arguments that directly support or undermine judgments and principles. For instance, when contemplating utilitarianism, the objection that utilitarianism seems in some circumstances to permit slavery or knowing punishment of the innocent should be considered (Varner 2012: 11). Second, there are structures for framing our deliberations that go beyond single arguments, which Rawls terms 'devices of representation' (Rawls 2005: 23-28). These typically provide special circumstances for principle selection, with the parameters of those circumstances set by the theorist's judgments regarding what is reasonable or rational. Rawls' original position is the best known example within analytical political philosophy. There are many more examples in contemporary work (Ackerman 1980; Gauthier 1986; Dworkin 2000: Ch. 2) and, arguably, older social contract theory (Hobbes, Locke) and ideal observer theory (Hume, Smith). Finally, there are 'background theories' (Daniels 1996: 22-23), which are drawn upon by both the direct arguments and the background theories, and themselves tested for intuitive appeal. For instance, if a theory of the separateness of persons were found compelling, it might be used both to undermine certain principles, as Rawls (1999: 23-24) seems to argue is true of utilitarianism. The various elements of wide reflective equilibrium are summarized in Table 1.

Element	*Scope*	*Role*	*Examples*
Judgments	Specific or general	Primary subject of equilibrium	Racial discrimination is wrong; all individuals have equal moral worth

Principles	General	Primary subject of equilibrium	The difference principle; equal moral worth principle
Direct arguments	Specific	Argumentation	Rawls' intuitive argument; the levelling down objection
Devices of representation	General	Argumentation	The original position; the ideal observer
Background theories	General	Argumentation	Theories of the person; social theory

Table 1. Elements of wide reflective equilibrium

In practice, it may not always be easy to distinguish the different elements, and it is not absolutely essential to do so. For instance, the table gives an example of a principle (equal moral worth principle) that is more or less a restatement of a judgment (all individuals have equal moral worth). As judgments and principles, qua judgments and principles, do not receive privileged epistemic status – '[o]ur "intuitions" are simply opinions: our philosophical theories are the same' (Lewis 1983: x; see also Freeman 2007: 33; Mandle 2009: 171-172) – there is no problem with the boundaries between them being fuzzy or overlapping.[2] Judgments and principles are only distinguished here as this is a familiar and often helpful way of arranging our thoughts. Likewise, and as indicated in the table, the direct arguments, devices of representation, and background theories are really just subsets of one big category of 'argumentation'. They do not need to be systematically separated as none has priority over any other. Some of the argumentation elements may even be absent in the creation of particular equilibria; for instance, the extended example in section 6 below does not refer to device of representation.

4. Challenges

As the most widespread approach to theory selection in moral and political philosophy, the method of reflective equilibrium has faced its share of critical attention. In this section I consider a couple of the more significant challenges.

A common complaint with the method is that it relies entirely on the quality of the judgments which form a central part of the equilibrium (Brandt 1979: 20; Williamson 2007: 244-246). Advocates of the method typically build their examples around highly plausible judgments, such as Rawls' convictions about the

[2] Welch even defends a radical version of reflective equilibrium in which 'there are no considered judgments to consider' (Welch 2014: 14).

wrongness of religious intolerance and racial subordination. But if someone starts with implausible or even repugnant judgments, there is, critics claim, nothing to stop the method from generating implausible conclusions. The point is put clearly by Thomas Kelly and Sarah McGrath (2010: 346-347):

> it is a good objection to a method if it turns out that impeccably following that method could lead one to views that are *unreasonable*. It follows from this that if beginning from all and only one's considered judgments, and from there achieving wide reflective equilibrium without making any "downstream" mistakes, is sufficient for impeccably executing the method of reflective equilibrium, then the method is not correct. The problem is that something might very well qualify as a considered judgment, when that notion is understood in anything like the way it is understood in the broadly Rawlsian tradition, and yet be utterly lacking in rational credibility.

This is illustrated with the observation that there is nothing to stop '[o]ne is morally required to occasionally kill randomly' from counting as a considered judgment. Kelly and McGrath therefore conclude that reflective equilibrium is an inadequate method.

This critique seems to be misdirected in several respects. First, Kelly and McGrath focus on considered judgments to such a degree that reflective equilibrium proper falls out of their picture entirely. They seem to take it as given that the final set of principles will simply be direct expressions of the initial considered judgments. While that may be more or less true of narrow reflective equilibrium, it is unlikely to be true of wide reflective equilibrium. Sustained consideration of competing principles, supporting arguments, devices of representation, and background theories is extremely likely to expunge judgments that are 'utterly lacking in rational credibility',[3] in which case the alleged problem does not arise.

Second, it is not clear that we have actually been shown a case in which 'impeccably following' the method of reflective equilibrium 'lead[s] one to views that are unreasonable'. In Kelly and McGrath's example, the random killing judgment is held *initially*. So it is not the case that the method 'leads' anyone to this judgment. Rather, they had the judgment to begin with. If there is a complaint to be had here, it is with the life history that has resulted in such an absurd judgment being formed.

Finally, I doubt that it actually is 'a good objection to a method if it turns out that impeccably following that method could lead one to views that are unreasonable'. Kelly and McGrath (2010: 327-328) support this claim with the following example:

> Suppose that, prior to embarking upon the systematic study of fruit flies, one held various baseless opinions about their nature. If one then devoted oneself

[3] Kelly and McGrath seem to concede a similar point regarding empirical sciences – see the lengthy quote given two paragraphs below.

> to the study of fruit flies, and impeccably followed the best scientific procedures we have for arriving at accurate views about their nature, we would expect those earlier baseless opinions to be filtered out or corrected at some stage in the inquiry. In the unlikely event that some of those opinions were among the views that one held after having impeccably following our best scientific methods, then, we submit, those beliefs would no longer be unreasonable ones to hold.

The conclusion may seem plausible here on account of misleading features of the case. In particular, the 'baseless opinions' are so sparsely described that we have no way of grasping whether they might be held reasonably or not. To really test the central underlying claim here that application of the scientific method, unlike reflective equilibrium, removes unreasonable beliefs, we should adjust the scientific baseless opinions, so that they are as vivid as their moral counterpart – the judgment that '[o]ne is morally required to occasionally kill randomly'. So suppose that the baseless beliefs about fruit flies are the following: fruit flies originate from specific acts of divine creation; these acts occurred within the last 10,000 years and are literally described in scripture; it is a matter of religious duty to disregard all countervailing evidence regarding the origins of fruit flies. I think it highly plausible that these views are unreasonable, and that 'devot[ing] oneself to the study of fruit flies, and impeccably follow[ing] the best scientific procedures we have for arriving at accurate views about their nature' does not stop these views from being unreasonable. The lesson to draw from this is that neither the method of reflective equilibrium nor the scientific method are *guaranteed* to rid people of unreasonable beliefs. But that doesn't change the fact that both are more likely than alternatives to provide individuals with reasonable beliefs, by exposing them to the most compelling evidence that is available in their respective fields.

This leads us to the second challenge. Several writers have claimed not that reflective equilibrium struggles with implausible idiosyncratic judgments, like the random killing judgment, but with the fact that our judgments are systematically undermined (Brandt 1979: 21-22; Hare 1981: 12). Peter Singer emphasizes that our moral judgments have largely arisen through an evolutionary process. For example, the common sense idea that we have stronger duties towards relatives can be explained on the basis that the corresponding genes 'are more likely to survive and spread among social mammals than genes that do not lead to preferences for one's relatives that are typically proportional to the proximity of the relationship' (Singer 2005: 334; see also Singer 1974). It is no surprise, then, that brain scans suggest that our moral judgments often do not seem to be informed by reason, but are rather an immediate emotional response (Singer 2005: 339-342). Individuals will stick to their judgment even where they end up rejecting the reasons they initially give for it (Singer 2005: 337-338). This modern scientific understanding of 'how we make moral judgments casts serious doubt on the method of reflective equilibrium' according to Singer (2005: 348):

> There is little point in constructing a moral theory designed to match considered moral judgments that themselves stem from our evolved responses to the situations in which we and our ancestors lived during the period of our evolution as social mammals, primates, and finally, human beings. We should, with our current powers of reasoning and our rapidly changing circumstances, be able to do better than that.

Suppose, for the sake of argument, that the evolutionary picture that Singer paints is correct. I would not see this as a threat to the method of reflective equilibrium. Singer is, in effect, presenting a background theory that should be considered when an individual is undergoing reflective equilibrium.[4] If the background theory is compelling, as I suspect it might be, that may cause individuals to treat their moral judgments differently, taking care to consider whether a judgment might amount to an evolved emotional response that should be set aside.

Singer (2005: 347) anticipates a response along these lines, and replies as follows:

> Admittedly, it is possible to interpret the model of reflective equilibrium so that it takes into account any grounds for objecting to our intuitions, including those that I have put forward. Norman Daniels has argued persuasively for this "wide" interpretation of reflective equilibrium. If the interpretation is truly wide enough to countenance the rejection of all our ordinary moral beliefs, then I have no objection to it. The price for avoiding the inbuilt conservatism of the narrow interpretation, however, is that reflective equilibrium ceases to be a distinctive method of doing normative ethics. Where previously there was a contrast between the method of reflective equilibrium and "foundationalist" attempts to build an ethical system outward from some indubitable starting point, now foundationalism simply becomes the limiting case of a wide reflective equilibrium.

Here Singer claims that reflective equilibrium would have to rely not just on the moderate, revisable foundationalism referred to earlier, but rather on a stronger 'special foundationalism' (Harman 2003: 415) that identifies certain ethical truths as unchallengeable. Were that true, it would certainly be the case that reflective equilibrium had been stripped of its distinctive features (in particular, mutual adjustment of judgments and principles). But it is not true. Singer says that the interpretation should be 'truly wide enough to countenance the rejection of all our ordinary moral beliefs'. Reflective equilibrium is this wide (Sandberg and Juth 2011: 222). However, Singer's conclusion implicitly assumes that countenancing the rejection of ordinary moral beliefs will result in (1) their wholesale rejection and (2) the adoption of some mysterious 'indubitable starting point', rather than a set

[4] Singer later seems to make this concession; see de Lazari-Radek and Singer 2012: 29-31.

of revised moral beliefs subject to reflective equilibrium's usual ongoing epistemic tests. Both of these assumptions are quite gratuitous. A more likely result of considering Singer's background theory is a reduction in the weight we are willing to assign to judgments that have a vividly personal quality, such as judgments favouring family members or judgments assigning special opprobrium to harms inflicted in a direct physical way, as these are likely to have evolutionary origins (Tersman 2008: 397-398). There may be a corresponding increase in the weight we are willing to assign to universal or impartial judgments, which have less (or possibly no) evolutionary baggage. Reforming our judgments in this way would not mean that 'the "data" that a sound moral theory is supposed to match have become so changeable that they can play, at best, a minor role in determining the final shape of the normative moral theory' (Singer 2005: 349). On the contrary, shifting judgments play a full role as part of a 'dynamic dialectical process' (Brink 2014: 688).

5. How to use the method of reflective equilibrium

In this section I suggest some steps in the process of reflective equilibrium.

The first step in reaching equilibrium is making considered judgments on the topic at hand. These are what I take to be the requirements for considered judgments (Rawls 1999: 42):

1) *No upset, fright, tiredness, or intoxication.* This may seem obvious, but there are plenty of cases where political theorists do their work when subject to personal distress, or to a deadline, or late at night, or (so I hear) over a glass of wine or two.
2) *No conflicts of interest.* Individual political theorists often would gain more under one set of principles rather than another. Even though there is effectively no chance that the principles are going to be put into effect, there may still be a psychological effect. This is actually a rather hard problem to get around – surely we cannot prohibit work on social justice, on the basis that any principle would be likely to have effects on theorists' incomes. Perhaps the best we can do is be aware of our possible subconscious biases, and exercise particular caution when rejecting principles that do not serve our self-interest.
3) *The ability to reach the correct decision.* This requires at least minimal standards of competency. It would be possible to reach a reflective equilibrium about a topic within political theory that one had never read anything about, but it is unlikely to have much epistemic value (see Scanlon 2014: 82).[5]

[5] It might be objected that this requirement seems incompatible with my rejection of the 'epistemic constraint' in section 2 above. This objection misses the importance of the distinction between constraints on the *contents* of judgments (such as the epistemic constraint) and constraints on the *circumstances* of judgments. The former type of constraint is otiose, as what it attempts to do (for instance, justification) is done more

4) *The desire to reach the correct decision.* The individual must be motivated to arrive at justified principles. People (almost?) invariably come to political theory with a set of preconceived ideas about politics. This is fine provided that the individual is open-minded, being willing to alter their views in response to arguments. The fact that one is on record defending a position should be no barrier to rejecting that position, even where this might prove inconvenient or embarrassing.[6]

In short, the first step is to make sure that you undertake the process of reflective equilibrium in the Rawlsian 'conditions favorable for deliberation and judgment in general'. **The conditions established in the first step must be maintained throughout the process.**

The second step is to draw up a list of the main contending principles on whatever topic you are considering. If you can think of any compelling new principles, these should also be added to the list. There is no specific number of principles that one should aim for, but as a general rule and time permitting, more is better. It should be remembered that, while it is usually impractical in a work of political theory of 5,000 or 10,000 words to discuss a large number of principles, there is no 'word limit' when it comes to considering principles prior to, or during, the actual writing process. Even if one only discusses two or three principles in detail in the final product, you may have considered and rejected many more during the process of reaching equilibrium. Presumably Rawls himself did – in the 'Presentation of Alternatives' section of *A Theory of Justice* he names over a dozen 'conceptions of justice', several of them containing multiple principles and one of them the extremely open ended 'list of prima facie principles (as appropriate)' (Rawls 1999, 107). While it would not be usual to provide such a lengthy list in writing, it is often useful to mention in passing your reasons for rejecting some of the principles that do not receive full discussion.

The third step is to begin reflective equilibrium in earnest. You go through each principle, checking its prescriptions against your judgments. Ask yourself: what are the central cases for my topic? And what are the hard cases for this principle? The literature is, of course, an invaluable resource for finding such cases, but you will also come up with your own. **Consider whether**

---

thoroughly by wide reflective equilibrium. The latter type of constraint is essential as its functions cannot be replicated by wide reflective equilibrium proper. For instance, a logical impossibility should be cleared from our judgments once we consider relevant background theory, provided we are reasoning in favourable circumstances. But the effects of unfavourable circumstances, such as being drunk or ignorant of relevant political theory, will not be cleared by reflective equilibrium, as the epistemic value of the process is fatally undercut by our adverse physical condition or inability to draw on relevant arguments, principles, and theories.

[6] A fifth step would be to expose oneself to a wide and representative range of non-philosophical experiences, in order to offset formative biases. While I am attracted to this proposal, it does go beyond the method of reflective equilibrium as usually conceived; DePaul 1993 treats it as part of the separate 'method of balance of refinement'.

you can accept the implications of the principle in each of these cases. It may be that initially the principle seems to have an unacceptable implication, but that on reflection you are willing to revise your judgment. This may particularly be the case where the principle is compelling in other cases. It may seem to *explain* why we think what we do in those cases, and *extend* in an appealing way to further, previously unconsidered cases. If you can accept a principle's implications, either right away or on reflection, then it would seem that this principle is worthy of further consideration. If you can't, you may set aside the principle for now, taking a note of the specific problems it faces. Repeat this procedure for each principle.

The fourth step is to bring in devices of representation and background theories (for example, the original position and a theory of the separateness of persons, respectively). The most important devices of representation and background theories relevant to the topic should be considered, with particular devices and theories chosen on the basis of our judgments, which may themselves be revised during the process, even in response to normative principles. It may be that you can find a device of representation that seems, at least on reflection, to capture reasonable constraints on theory selection. It may even be that you have more confidence in it than you have in any principle. For instance, I personally find the difference principle less plausible than the original position from which Rawls controversially (Harsanyi 1975; Hare 1975: 102-107) derives it. In such cases, you may decide to focus on the principles chosen from the circumstances specified by the judgment-endorsed device, though you are still free to directly check the chosen principles against considered judgments (Mandle 2009: 40). Background theories have a similar, though less dramatic role, guiding principle selection but not outright replacing direct reference to judgments. Plausible background theories are used at this point to assess principles, with a particular focus on the principles found appealing in stage three. Devices of representation are also tested by background theories. For instance, if we accepted Sandel's (1982: ch. 1) claim that the original position assumes that the self is prior to values, we might reject that device of representation as incompatible with our favoured non-moral background theory even if it were compatible with our normative judgments (Gaus 1996: 105).

Having considered principles, devices of representation, and background theories, **the fifth step is to review this process. Now you know the specific challenges faced by the various principles, are there any revisions to these principles worth considering? Or do any entirely new principles now come to mind?** If so, the third and fourth step should be repeated for these principles. If new or revised principles keep arising, many iterations of the third and fourth step may be necessary. The same repetition applies where revised or new devices of representation and background theories arise. Likewise, if you have *not* found any principle that you find acceptable in their implications, the third and fourth steps should be repeated. It may, however, be acceptable to limit the level of repetition due to time constraints. We do not all have months or years of philosophical

contemplation available to us! If all the steps are followed, the method will yield dividends even if the fifth step is attenuated, as may be necessary if writing a student essay, for example.

The sixth step is to establish priority rules. This applies only where you have accepted multiple principles that may come into conflict with each other. Where you have such principles you need to consider cases of conflict, and decide how much importance each principle has in them. It may be that one principle seems so important that it should have absolute or 'lexical' priority over another. Alternatively, the principles may seem to have similar importance, in which case some kind of weighting should be decided. It could even be found that there are 'incompatible but equally justified overall accounts of the subject, thus supporting a kind of pluralism about the subject' (Scanlon 2014: 78-79).

The seventh step is the conclusion of the process, insofar as it has one. By this point you should have found agreement between principles and judgments – or otherwise concluded that this is impossible as there are no acceptable principles! Either way, **your findings are only ever provisional, and should be considered permanently open to revision.**

6. An example

I will now work through an example of reflective equilibrium on the topic of distributive justice, using the above step-by-step guide and my own considered judgments. I can obviously give only the scantest indication of my reasoning here, summarizing years of work in a few paragraphs. It is likely, furthermore, that the reader will disagree with me at numerous points. The example should nevertheless illustrate one way of reaching reflective equilibrium.

The first step is to make sure that my judgements are considered. As I write this, it is 9.32 am, I had a good night's sleep, I am aware of the danger of conflicts of interest when discussing the societal allocation of goods and am willing to counteract any resulting bias, and I have the motivation and desire to reach the correct decision. So it seems that, right now, I am making my judgments in suitable conditions. But as this test must be taken each time you use the method of reflective equilibrium, it must be repeated many times – indeed, many thousands of times in my case!

For the second step I have to draw up a list of the main principles within this topic. Here's my list:

- The principle of utility
- Rawls' two principles of justice
- Equality of outcome
- Luck egalitarian principles (Arneson 1989; Cohen 1989)

- Democratic egalitarian principles (Anderson 1999)
- The principle of priority (Parfit 2000)
- The principle of sufficiency (Frankfurt 1987)
- Right libertarian principles (Nozick 1974: Ch. 7)
- Left libertarian principles (Steiner 1994)
- The benefiting principle (Butt 2007)
- The principle of need
- The principle of desert
- Communitarian principles (Sandel 1982)
- Contractarian principles (Gauthier 1986)
- Egoist principles

The list is eclectic, and by design – the point at this stage is to avoid missing anything important, not to construct the most elegant inventory possible. Even so, other people's lists would no doubt contain additional principles.

With the third step I begin the reflective process by testing the principles and judgments against each other. Many principles can be set aside quite quickly. I find nothing of merit in 'free for all' egoist principles, and view the results of contractarian principles for people with low bargaining power utterly unacceptable, for instance. I do, by contrast, feel the pull of the principle of sufficiency, as I am very concerned by those who are very badly off in absolute terms. But I do not accept its implication that those who are just below the threshold of 'having enough' get absolute priority over those marginally above the threshold, who are only slightly better off (Arneson 2006: 28). I therefore conclude that the principle of priority better accommodates my concern with the absolutely badly off. Similarly, I am attracted to equality of outcome and democratic equality, as I am also concerned about inequality. But I am unhappy with democratic equality's implication that large unchosen inequalities do not matter as long as individuals have equal social standing, and equality of outcome's implication that, where some squander their equal share of resources, for instance by deliberately developing 'expensive tastes' (Dworkin 2000: 48-59), they should be 'compensated' to restore equality, at society's expense. I find that luck egalitarianism, which avoids such problems, fits with my judgments here better, but it has its own apparently objectionable implication that those who make bad choices that leave them in severe disadvantage will be 'abandoned' (Anderson 1999: 295-296). Outcome egalitarianism has no such implication. So at the end of third stage I have a provisional endorsement of prioritarianism, and an interest in egalitarianism that I am not yet convinced is well expressed in any principle.

The fourth step sees the introduction of background theories (I set aside devices of representation). I will mention only one line of thought here, to illustrate how background theories might help us arbitrate between political principles. Some critics of luck egalitarianism have claimed (1) that it assumes that metaphysical libertarianism (the theory that free, non-causally determined human action is

possible) is true, and (2) that metaphysical libertarianism is false (Scheffler 2003: 17-19). Were this true, I would have a background theory-based reason to reject luck egalitarianism in favour of outcome egalitarianism or democratic equality. However, on reflection I find reasons for rejecting both claims. While (2) is possible, we do not have adequate grounds for assuming this to be the case; political theorists would do better to proceed under the assumption that any of the main theories of free will (including sceptical views such as hard determinism) might be correct (call this the 'thin theory'). Regarding (1), the standard, Arneson-Cohen construal of luck egalitarianism does not after all assume any theory of free will, but is instead responsive to the morals and metaphysics of responsibility, in the sense that what counts as 'chosen' (and therefore as potential justification for inequality) depends on the best philosophical account. If metaphysical libertarianism is false, this just means that one way in which choice might have arise can't actually happen. Luck egalitarianism would even be compatible with there being no way for true choice to arise. In that case, no inequality would be justified, a point which mitigates the 'abandonment objection' to luck egalitarianism mentioned in the previous paragraph (Knight 2015: 132-134). So luck egalitarianism is in fact admirably responsive to what I take to the most plausible background theory about free will, which is the thin theory (Knight 2009: ch. 5). Outcome egalitarianism and democratic equality are not responsive in this way, however, as they make the same prescriptions whether metaphysical libertarianism is true or hard determinism is true. This seems a significant flaw to me as, in my judgment, where a person has prima facie brought some hardship upon herself, we have more reason to assist her if her action were not a true exercise of free will, and less reason to assist her if action were a true exercise of free will. As luck egalitarianism seems to accommodate the most plausible background theory better than rival egalitarian theories, I accept it as part of my wide reflective equilibrium.

For simplicity, I leave aside the fifth step. This brings us to the sixth step. I have found luck egalitarianism and prioritarianism to be in accord with my judgments. Now we need to decide on a rule to regulate conflicts between these principles. It seems that neither the 'eliminate involuntary disadvantage' (Cohen 1989: 916) goal of luck egalitarianism, nor prioritarianism's concern with increasing absolute advantage (in particular, that of the worst off), should be assigned lexical priority, as I would be willing to give up a small improvement in either of these dimensions for a large improvement in the other. So my conception of justice of justice is a version of 'responsibility-catering prioritarianism' (Arneson 1999; see also Knight 2009: ch. 6), where luck egalitarianism and prioritarianism are balanced against each other. Exactly what weighting, however, to give each of these principles is a rather tricky question, to be tested through considering a large number of cases, which I cannot do here.

Suppose, though, that I find a favoured weighting, and reach the seventh and final step. Even then I can't assume that weighting, or even the selection of principles,

to be settled for all time, as we can reconsider any aspect of the process at any point. As Rawls (2005: 97) cautions, '[t]he struggle for reflective equilibrium continues indefinitely'.

References

B. Ackerman, 1980, *Social Justice in the Liberal State* (New Haven: Yale University Press).

E. S. Anderson, 1999, 'What is the point of equality?', *Ethics* 109, 287-337.

R. J. Arneson, 1989, 'Equality and equal opportunity for welfare', *Philosophical Studies* 56, 77-93.

R. J. Arneson, 1999, 'Equality of opportunity for welfare defended and recanted', *Journal of Political Philosophy,* 7, 488-97.

R. J. Arneson, 2006, 'Distributive justice and basic capability equality' in A. Kaufman (ed.), *Capabilities Equality* (Abingdon: Routledge), pp. 17-43.

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

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

D. O. Brink, 2014, 'Principles and intuitions in ethics', *Ethics* 124, 665-695.

D. Butt, 2007, 'On benefiting from injustice', *Canadian Journal of Philosophy* 37, 129-52.

G. A. Cohen, 1989, 'On the currency of egalitarian justice', *Ethics* 99, 906-44.

R. C. Cummins, 1998, 'Reflection on reflective equilibrium' in M. R. DePaul and W. Ramsey (eds.) Rethinking Intuition (Lanham: Rowman and Littlefield), pp. 113-128.

N. Daniels, 1996, *Justice and Justification* (Cambridge University Press).

K. de Lazari-Radek and P. Singer, 2012, 'The objectivity of ethics and the unity of practical reason', *Ethics* 123, 9-31.

M. R. DePaul, 1986, 'Reflective equilibrium and foundationalism', *American Philosophical Quarterly* 23, 59-69.

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

R. Dworkin, 2000, *Sovereign Virtue: The Theory and Practice of Equality* (Cambridge, MA: Harvard University Press).

R. Ebertz, 1993, 'Is reflective equilibrium a coherentist model?', Canadian Journal of Philosophy 23, 193-214.

H. Frankfurt, 1987, 'Equality as a moral ideal', *Ethics* 98, 21-43.

S. Freeman, 2007, *Rawls* (Abingdon: Routledge).

G. Gaus, 1996, *Justificatory Liberalism* (Oxford University Press).

D. Gauthier, 1986, *Morals by Agreement* (Oxford University Press).

N. Goodman, 1965, *Fact, Fiction, and Forecast* (Indianapolis: Bobbs-Merrill).

R. M. Hare, 1975, 'Rawls' theory of justice' in N. Daniels (ed.), *Reading Rawls* (Oxford: Blackwell), pp. 81-107.

R. M. Hare, 1981, *Moral Thinking: Its Levels, Method, and Point* (Oxford University Press).

G. Harman, 2003, 'Three trends in moral and political philosophy', *Journal of Value Inquiry* 37, 415-425.

J. C. Harsanyi, 1975, 'Can the maximin principle serve as a basis for morality?', *American Political Science Review* 69, 594-606.

B. Hooker, 2012, 'Theory vs anti-theory in ethics' in U. Heuer and G. Lang, (eds.) *Luck, Value, and Commitment: Themes from the Moral Philosophy of Bernard Williams* (Oxford University Press), pp. 19-40.

T. Kelly and S. McGrath, 2010, 'Is reflective equilibrium enough?', *Philosophical Perspectives* 24, 325-359.

C. Knight, 2006, 'The method of reflective equilibrium', *Philosophical Papers* 35, 209-225.

C. Knight, 2009, *Luck Egalitarianism: Equality, Responsibility, and Justice* (Edinburgh University Press).

C. Knight, 2015, 'Abandoning the abandonment objection: luck egalitarian arguments for public insurance', *Res Publica* 21, 119-135.

D. Lewis, 1983, *Philosophical Papers,* volume 1 (Oxford University Press).

S. Maffetone, 2010, *Rawls: An Introduction* (Cambridge: Polity).

J. Mandle, 2009, *Rawls's* A Theory of Justice: *An Introduction* (Cambridge University Press).

D. McDermott, 2008, 'Analytical political philosophy' in D. Leopold and M. Stears (eds.) *Political Theory: Methods and Approaches* (Oxford University Press), pp. 11-28.

R. Nozick, 1974, *Anarchy, State and Utopia* (Oxford: Blackwell).

D. Parfit, 2000, 'Equality or priority?' in M. Clayton and A. Williams (eds) *The Ideal of Equality* (Basingstoke: Palgrave), pp. 81-125.

J. Rawls, 1975, 'The independence of moral theory', *Proceedings and Addresses of the American Philosophical Association* 48, 5-22.

J. Rawls, 1999, *A Theory of Justice,* revised edition (Oxford University Press).

J. Rawls, 2005, *Political Liberalism,* expanded edition (New York: Columbia University Press).

J. Sandberg and N. Juth, 2011, 'Ethics and intuitions: a reply to Singer', *Journal of Ethics* 15, 209-226.

M. Sandel, 1982, *Liberalism and the Limits of Justice* (Cambridge University Press).

T. Scanlon, 2003, 'Rawls on Justification' in S. Freeman (ed.) *The Cambridge Companion to Rawls* (Cambridge University Press).

T. Scanlon, 2014, *Being Realistic About Reasons* (Oxford University Press).

S. Scheffler, 2003, 'What is egalitarianism?', *Philosophy and Public Affairs* 31, 5-39.

P. Singer, 1974, 'Sidgwick and reflective equilibrium', *The Monist* 58, 490-517.

P. Singer, 2005, 'Ethics and intuitions', *Journal of Ethics* 9, 331-352.

W. Sinnott-Armstrong, L. Young and F. Cushman, 2010, 'Moral intuitions' in J. M. Doris and The Moral Psychology Research Group, *The Moral Psychology Handbook* (Oxford University Press).

F. Tersman, 2008, 'The reliability of moral intuitions', *Australasian Journal of Philosophy* 86, 389-405.

H. Steiner, 1994, *An Essay on Rights* (Oxford: Blackwell).

G. E. Varner, 2012, *Personhood, Ethics, and Animal Cognition.* Oxford: Oxford University Press.

Walden, 2013, 'In defense of reflective equilibrium', *Philosophical Studies* 166, 243-256.

J. R. Welch, 2014, *Moral Strata: Another Approach to Reflective Equilibrium* (Dordrecht: Springer).

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

J. Wolff, 2013, 'Analytic political philosophy' in M. Beaney (ed.) *Oxford Handbook of the History of Analytic Philosophy* (Oxford University Press), pp. 795-822.