The Method of Reflective Equilibrium: Wide, Radical, Fallible, Plausible[*]

Carl Knight

1. Introduction

How can moral principles be justified? If such principles could be convincingly derived from morally-neutral facts we would have a ready answer. But no such derivation is possible.[1] In the absence of such a possibility we have little choice but to justify moral principles by some kind of appeal to human judgment.[2] But what kind of judgments should we use, and how should we use them?

John Rawls' answer to the first part of this question was *considered judgments* – those judgments made under idealized circumstances and without errors. Famously, and controversially, his answer to the second part of the question was *the method of reflective equilibrium,* according to which considered judgments and moral principles can be matched by a process of philosophical reconciliation.[3] This article will clarify, modify and defend the role of considered judgments and the method of reflective equilibrium in several ways and attempt to settle some of the areas of controversy.

The paper begins by setting out the character of considered judgments and the method of reflective equilibrium in a little more detail, before considering the viability of the independence constraint on wide equilibria and the confidence constraint on judgments. A radical conception of the method is then set out that emphasizes the role of individuals undergoing the reflective process in tailoring their own principles and the epistemological value of considering other persons' beliefs and undergoing a balanced range of potentially formative activities. It is subsequently maintained, contrary to a common view, that the method of reflective equilibrium is not an outright or even primarily coherentist method but rather represents a balance of coherentism and fallibilist foundationalism. This all sets the stage for assessing major criticisms of both moral judgment-based approaches to moral justification in general and the method of reflective equilibrium in particular. In response to the first branch of criticism, considered judgments are shown to be both sufficiently initially credible (on the basis of internal considerations) and satisfactorily cleared of bias and prejudice (due to the reference to background theories and other persons' beliefs, and the range of activities undergone). The second branch of criticism has it that that the method of reflective equilibrium fails to determine theory and principle selection; but this complaint assumes that there is some kind of judgment-independent way of arbitrating such matters, which there is not. I conclude that an appropriately modified version of the method of reflective equilibrium is a fair and plausible approach to moral justification.

2. Two Constraints Considered

Rawls defines considered moral judgments 'as those judgments in which our moral capacities are most likely to be displayed without distortion'.[4] We start by presuming the 'the ability, the opportunity, and the desire to reach a correct decision (or at least, not the desire not to)'.[5] Individuals must satisfy a minimal level of competency and be motivated to identify true moral principles. Distortions in our judgments are more likely to be present where we are partial because we may profit or lose out, or where we are upset, frightened, tired or intoxicated.[6] Even where these conditions are removed, however, distortions may occur. It is therefore necessary for us to identify the character of distortions themselves. Errors of reasoning, such as false inferences and logically inconsistent uses of concepts (and the beliefs from which they are constructed), are one type of distortion to be avoided. Further, we must take as a precondition for a considered judgment that it is not founded on empirical error.[7] If a judgment is logically or empirically false it cannot be a considered judgment.

The requirements already described are externalist constraints on the method, though fairly unobjectionable ones. There is a further requirement that is conventionally thought necessary for judgments to count as considered. The *confidence constraint* demands that the individual has confidence in her judgments, displaying little or no doubt that they are superior to the alternatives.[8] T. M. Scanlon notes that there is nothing wrong with the aim of an equilibrium which accounts for judgments in which we have little confidence, but accepts that accounting only for those in which we are confident may be a good starting point.[9] I doubt that such a limited approach is even a sensible place to start.

Suppose that there are two principles, P and Q, each of which implies a set of ten judgments, P' and Q' respectively, on the same ten cases. Also suppose that I hold ten out of ten of the P's with little confidence, one of the ten Q's with complete confidence, and have no confidence in the remaining nine Q's. I now apply the confidence constraint when I start the reflective process and, upon examining the two principles, find Q to be the better fit as it is the only one of the two to coincide with my considered judgments at all.

It seems at least prima facie plausible that P better captures my considered judgments on this range of issues, for in nine out of ten cases it outperforms Q. It is possible that my commitment to the one endorsed Q' is of such a magnitude that that one hefty benefit of Q outweighs its numerous minor disadvantages. But equally, our commitment to the endorsed Q', solid as it may appear, could be outweighed by the sheer volume of endorsed P's. To treat them as though they do not count at all is bizarre. We should not deny ourselves the opportunity to undergo the full reflective process, where we weigh various considerations against one another, allowing for the depths of our beliefs on every relevant issue. Those judgments in which we are least confident are, of course, 'more likely to be jettisoned in the process of reaching an equilibrium'.[10] But whether such judgments are rejected or not ought to be a matter for the individual; it cannot be arbitrarily decided in advance. I am moved, then, to drop the confidence constraint. Any judgment that meets the earlier requirements is taken to be considered.

The distortion clearing procedure embodies a certain critical standard that is already publicly acknowledged – people readily reject judgments known to be based on bad information, faulty deductions and personal interest.[11] Unfortunately it is also rather undemanding, often leaving many mutually contradictory judgments in place. This is where the method of reflective equilibrium comes into play. Rawls writes that '[w]hen a person is presented with an intuitively appealing sense of justice (one, say, which embodies various reasonable and attractive presumptions), he may well revise his judgments to conform to its principles even though the theory does not fit his existing judgments exactly'.[12] The idea is that the range of diverse remaining judgments can be brought, by philosophical means, into the equilibrium of a theory and its principles.

This core idea can be fleshed out in two contrasting ways. A *narrow reflective equilibrium* consists of an individual's set of considered moral judgments, and a set of general moral principles that systematizes the judgments in parsimonious fashion. The principles create *coherence* in the individual's set of judgments, unifying and explaining the views we hold and extending the range of judgments beyond those areas in which the individual initially had confidence.[13] Just as a grammar may be devised to describe a person's syntactic competency, so narrow reflective equilibrium may establish moral principles that characterize a person's moral competency or 'sensibility'.[14] *Wide reflective equilibrium,* as the name suggests, casts the justificatory net significantly wider. Several moral conceptions are to be considered, and philosophical arguments advanced to show the strengths and weaknesses of each conception.[15] Various *background theories,* such as theories of the person, theories about the role of morality in society, and general social theory, should also be taken into account.[16] In order for wide reflective equilibrium to be reached, each moral principle under consideration is to be assessed not only by direct reference to the individual's considered moral judgments on that particular question, but also by reference to relevant background theories. The principles shown (if necessary, by argument) to be associated with the most plausible background theories may consequently be favoured, regardless of their fit with the initial considered moral judgments.

It is generally accepted that the method of wide reflective equilibrium is an advance on its narrow counterpart (and any other approach to moral justification that only looks at moral judgments). Principles held in wide equilibrium cohere not merely with the individual's moral judgments, duly systematized, in the immediate area, but with her entire system of considered judgments over a massive area of thought, after it has undergone the most thorough examination. As Michael DePaul notes, 'it would seem irrational for a person to go on believing the moral theory that is fully coherent with his moral beliefs if he were to realize that it does not cohere with non moral beliefs … which the person finds more compelling than any of his moral beliefs'.[17]

There is, however, an immediate concern. How do we know that the favoured background theories, and their corresponding moral principles, are favoured independently of those principles' match with considered moral judgments? If the choice of these theories and principles is not independent in this way there is no justificatory advantage over narrow reflective equilibrium, for the reference to background theory has been rendered redundant: 'How, it will be asked, can we have any confidence in moral principles justified by appeal to background theories that are, in large part, justified by the very same principles they are supposed to justify?'.[18] In the face of this apparent circularity, Norman Daniels proposes an *independence constraint*, which requires that the background theories are supported by an independent set of considered judgments.[19] We might call this set, which establishes the plausibility of background theories, *background judgments.* For example, Rawls' contract device may add justificatory force to his theory if it embodies certain background theories that receive intuitive support that is quite independent of the intuitive support his principles directly receive.[20]

D. W. Haslett is unhappy with Daniels' response:

> Deliberately not to take some of our considered judgments into account in arriving at a narrow reflective equilibrium would appear to have no other purpose than that of keeping them held back in reserve so that they can be used for the first time in arriving at a wide reflective equilibrium, thus allowing us to claim

> that the independence constraint has been satisfied. But if the independence constraint can only be met by thus artificially limiting the number of considered moral judgments we use in arriving at a narrow reflective equilibrium, then surely meeting it does not very satisfactorily enable us to avoid the circularity objection.[21]

The complaint here against the independence constraint is that, as it requires that less than all considered judgments are included in the initial narrow reflective equilibrium, it permits the excluded judgments to be reintroduced later, distorting the process. But the complaint fails to grasp the purpose and effects of the constraint. The constraint just ensures that none of the considered moral judgments are also later readmitted as background judgments. Where there is the potential for such double-counting there is no question of holding back considered moral judgments so they can be used as background judgments – they will be used as considered moral judgments, and the background theories will either have to be supported by other judgments or not be supported at all. There is, in a certain sense, an artificial limit on the kinds of considered moral judgments used in reaching the initial narrow reflective equilibrium, but that limit prevents the distortion associated with double-counting and creates no new distortion.

3. Radical Reflection

Having accepted the independence constraint, and rejected the confidence constraint, I will now set out the scope of the reflective process in broader terms. I will begin by asking two questions, my answers to which suggest a radical interpretation of the method of reflective equilibrium.

The first question is this: *Which moral principles* are available for selection? In *A Theory of Justice,* Rawls famously opts to present just a couple of alternative theories, one of which is to be accepted wholesale.[22] But we would ideally expect people to view the body of moral theory *as a whole*, along with corresponding background theory, placing each existing theory on an equal footing. Further, I would insist that individuals may make their modifications to their judgments on the basis of the parts of each theory which, on reflection, they find to be most attractive. Conversely, they may tailor their own hybrid theories, in accordance with the particular intuitive weightings they place on each part of each theory. This allows a person to revise some of her judgments to comport with a theory, where that theory plausibly explains most of her other judgments, without making the unreasonably strong demand that she revise all of her nonconforming judgments, including those which are deeply held, and perhaps even reinforced by reflection. The approach described should both better inform its participants – for they will be aware of a wider range of theories and principles – and be more sensitive to the diverse and complex systems of considered judgments that may be held during the reflective process – for while justice as fairness may reflect my considered judgments more closely than utilitarianism or any other classical theory, there can be no guarantee that it reflects them better than any further, as yet unproposed, theory. This should help to protect the method of reflective equilibrium from the charge that it favours parsimony of principles even at the expense of moral knowledge.[23]

The second question is this: How should persons undergoing the reflective process view *other person's judgments?* In *A Theory of Justice,* Rawls writes that 'for the purposes of this book, the views of the reader and the author are the only ones that count'.[24] But we can surely only have faith in our own judgments where we have looked at those of others. The point is put forcefully by David Miller:

> can we decide whether a judgment is considered simply by scrutinizing it in solipsistic fashion, relying only on internal evidence to establish how much confidence we should place in it . . . ? It is surely of the greatest relevance to see whether the judgments we make are shared by those around us, and if they are not, to try to discover what lies behind the disagreement.[25]

I am unsure how effective this is as a criticism of Rawls. The very next sentence of *A Theory of Justice* reads '[t]he opinions of others are used only to clear our own heads'.[26] Although this comment is left unexplained, it seems likely that Rawls has some interest in alternative views. But whatever Rawls stance may be, it seems clear that we should take the moral judgments of other people seriously when arriving at our own moral judgments. The method of reflective equilibrium should help an individual to work out the characteristics and effects of moral principles, how these relate to what else she knows, how her particular status might influence how she sees them, and how much of this she can collectedly consider to be relevant to the justification of principles. Evidently, considering the beliefs of other people is entirely consistent with and (given the need to find how we might be influenced by our particular social standing) necessary for this process. Even where we are only devising principles for our own society the requirement extends to having an understanding of how views vary between societies, for this understanding may suggest how our social institutions have shaped our moral views.

But while the individual must consider others' views, the final decision can only be hers if we wish to maintain the critical standard which we must surely attach to moral theory. It may be instructive to note some differences with the position occupied by communitarians. For Michael Walzer, 'in matters of morality, argument simply is the appeal to common meaning'.[27] But social meaning is very likely to be contingent on factual error, defective inference, self-interest, historical accident and brute strength. A communitarian approach to justification therefore robs moral theory of its crucial critical function.[28] We can acknowledge that the agent is socially conditioned while accepting that any attempt to negate or override rather than account for this is liable instead to obscure its effects.

These considerations dovetail with DePaul's *radical interpretation of the method of reflective equilibrium.* As DePaul describes them, the conservative interpretation permits revision of judgments solely on the basis of incoherencies in the individual's set of considered judgments, and these revisions are determined by the individual's levels of *initial* commitment. Once the initial considered judgments are defined, 'anyone with the relevant rational capacities' or 'perhaps a machine' could carry out all the justificatory work required.[29] The radical interpretation, by contrast, grants far greater powers of revision, with judgments being altered wherever reflection suggests, regardless of initial commitment.[30] Furthermore, it recognizes the role of 'experiences other than the consideration of philosophical arguments … as playing a significant and legitimate role in the development of a coherent moral view'.[31]

DePaul argues, I believe rightly, that the radical conception is much the more defensible.[32] The assumption of the conservative method appears to be 'that when we begin moral enquiry we already possess as much of the truth about morality as we ever will'.[33] We cannot be satisfied with a mere 'tidying up' of our incoherent pre-philosophical moral judgments in order to bring out their supposed underlying insights. This would be to deny to ourselves much of the critical power the method of reflective equilibrium apparently offered. We must account for the possibility that philosophical engagement – for example, critical assessment of various moral principles – will convince us that our initial moral judgments were outright mistaken.

Even if we accept that philosophy may legitimately lead us to accept one moral view rather than another, it might seem prima facie strange to claim that such a view could be justified, even to me, by my experience of, say, watching a film devoid of philosophical content. But at the same time, many other experiences will have shaped my moral views, and it seems only appropriate to compensate for any biases in these experiences that I can detect. If the film depicted the daily lives and views of black working class women and I am a white middle class man it may well be an appropriate corrective to the limitations of my prior formative experiences. Thus each individual undergoing the radical reflective process is required to engage in all (available) activities that may offset any of her formative biases.[34]

The move to radical reflective equilibrium, like the earlier one concerning the consideration of other persons' views, is intended to boost the epistemological standing of the judgments of the person undergoing reflective equilibrium. In general we have more justification for believing those beliefs that are formed with knowledge of relevant background facts and after experiencing a full variety of activities. The next section will focus on the epistemological status of the method of reflective equilibrium itself.

4. Coherentism, Foundationalism, Fallibilism

One of the oldest problems in epistemology is that of infinite justificatory regress. We conventionally seek to justify beliefs in an inferential and linear fashion. Such a chain of justification goes like this: belief A is justified by belief B, which is justified by C, which is justified by D, which is justified by E, and so on ad infinitum. The chain is *inferential* as each belief is justified by another belief. The chain is *linear* because none of the justified beliefs can justify any of the antecedent beliefs. The problem is that the inferences must come to a premature end as the human brain cannot conceptualize an infinite number of beliefs. The chain of justification is therefore broken and, seemingly, none of our beliefs can be justified. There are two obvious ways out: we may maintain that all justification is linear but deny that all justification is inferential, in which case we are *foundationalists.* Alternatively, we may maintain that all justification is inferential but deny that all justification is linear, in which case we are *coherentists*.[35]

Wide reflective equilibrium is conventionally taken to be a coherentist approach to justification.[36] Dropping the linearity requirement means that the chain may loop back on itself; in our earlier example we might get out of the vicious regress by saying that E is justified by A. This invites the complaint of vicious circularity, for each belief is in the end justified by itself. More recently, however, some have suggested that wide reflective equilibrium may be consistent with foundationalism.[37] In that case we might prefer to say that E is justified in a way that cannot be inferred from any other belief, or at least that cannot be inferred from any other *moral* belief. Unless E can be justified on the basis of definitional truths, which is unlikely in the case of the substantive moral beliefs in which we are interested, this invites the claim that we are being told to believe E for no good reason at all.

There are only dubious grounds for favouring a strong coherentist interpretation. Daniels writes that wide reflective equilibrium cannot be foundationalist as it permits 'drastic *theory-based* revisions of moral judgments'.[38] This may or may not mark a noteworthy break from traditional foundationalism, which in some of its variants permits revision of initial intuitions.[39] But we can certainly imagine a foundationalist theory that permitted very extensive revision of considered judgments in the light of philosophical reflection. To be foundationalist, a theory need not subscribe to what Gilbert Harman calls 'special foundationalism' – the view that foundational beliefs are 'either self-evident or directly justified by experience'.[40] Rather, it need only rule out non-linear justification and rule in non-inferential justification. Indeed, this latter foundationalist ruling may seem more plausible where the self-justifying beliefs have been arrived at after considering all relevant alternatives and all corresponding background theories. (The background theories would not, of course, serve the same justificatory role as they do in reflective equilibrium). In other words, although we cannot infer our most basic moral beliefs from anything else, they receive some support from various non-moral beliefs; the justificatory gap is narrowed.

David Brink claims that reflective equilibrium is not foundationalist as its considered judgments are justified by reference to their coherence with other beliefs.[41]

However, the crux of the matter is the justificatory role of beliefs themselves. Suppose that there are two mutually contradictory but internally coherent packages, each consisting of a set of moral principles, considered moral judgments, background theories and background judgments. Further, suppose that, for each of these sets, acceptance of the judgments implies acceptance of the theories and principles, that there are no relevant considerations (plausible supporting judgments, theories, etc.) lying outside of the packages, and that neither package amounts to a set of definitional truths – they are morally substantive. Finally, suppose that, after full consideration, my judgments reach equilibrium, and I accept the first package as true in every detail and therefore believe the second to be false. There is no question, it can be agreed, that the outcome of the process is that the first package is to be favoured. The reason for this is simply that, under the relevant conditions, I *believe in* the first package. After all, someone else in my place could have gone through the same procedure yet arrive at the contrary conclusion. We cannot explain this outcome in coherentist terms for, in the specified case, the coherentist chains of justification start and end inside each of the packages. The first package is justified to the agent by the self-justifying (i.e. foundationalist) reason that the individual favours his favoured beliefs. In this case, where there is only one agent (and therefore only one equilibrium), agent and propositional justification are one and the same.[42]

It perhaps seems that I have laboured an elementary point with little relevance for the present question. Reflective equilibrium does assume this minimal foundationalist requirement, it might be said, but all the hard work is nevertheless done in a coherentist manner. This is not so. Roger Ebertz describes an uncontroversial description of the test of principles offered by wide reflective equilibrium: it is ‘the test of whether they fit the considered judgments we are committed to *at that point in the reflective process.* If they do not they are not justified.'[43] As Ebertz concludes, it is clear that our considered judgments have a justificatory function that is quite distinct from requirements of coherence. Judgments, principles and background theories can be perfectly coherent, but if those judgments are not actually held by the individual undergoing the reflective process then that coherence is irrelevant. The minimal foundationalist requirement that our actually favoured judgments must set the point of equilibrium is acknowledged throughout the process.

None of this is intended to refute the role coherence plays in reaching reflective equilibrium. The earlier discussion showed how nonlinear justification can play a role in wide reflective equilibrium. The variant of the method of reflective equilibrium outlined here uses a combination of coherentist and foundationalist methods of justification. We may justify E either by direct appeal to a self-justifying considered judgment, by reference to background theories (assuming that they there are corresponding self-justifying background judgments and that the independence constraint is not breached), by looping our chain of justification back to A, or, as is more likely, by more than one method. The foundationalist method is perhaps the more fundamental. There would be little justification for the belief set A-E regardless of internal chains of reasoning if we did not believe any of its constituent beliefs or supporting theories, whereas it may well be justified if we believed each constituent belief and supporting theory but there were no internal chains. But it would be overstating the case to say that it is outright foundationalist. We can certainly envisage circumstances where the strength of inferential, nonlinear justification – where, for instance, a set *explains* our views to us better than a mishmash of equally intuitively attractive but mutually unconnected beliefs – may tip the balance even where our beliefs are initially – that is, before we have taken coherence considerations into account – tilted in the opposite direction. Furthermore, insofar as the method is foundationalist it is also fallibilistic, for the foundations (the considered judgments) are uncertain and subject to revision: we use them only until we have good reason to replace them.[44]

Is this fallibilism a problem? As noted, the method of radical reflective equilibrium states that a change in our considered judgments can change the status of principles as justified. Taking this acceptance together with the observation that people's beliefs can change while the relevant facts do not, David Copp concludes '[t]hat this undermines the plausibility of taking our considered moral judgments to constitute a standard of justification, for it shows the putative standard to be a drifting one.' Judgments can indeed be changed in this way. Unfortunately Copp does not explain how or why the 'drifting' undercuts the 'putative standard.'[45] To be sure, if there were one conception of morality that was indisputably superior to all its rivals we would prefer that persons' judgments did not drift from it. But as there is no such conception it seems only reasonable to allow persons to change their minds about morality following, say, consideration of philosophical arguments or new experiences.[46] What he might have in mind is the moral realist's worry that such drifting demonstrates that equilibria *must* fail to realize moral truth, for, on the realist account, a moral truth exists independently of beliefs about it. If such disquiet is what Copp has in mind it is misplaced. If the realist is wrong then we can obviously forget about his concern. But even if realism is true there is still no problem for the method of reflective equilibrium. The key point is that the reflective process can be interpreted by a realist as a way of *justifying* beliefs, not as a way of arriving at truth itself. Any of the beliefs that are in equilibrium may be false. But it remains reasonable to believe them as long as there are no more plausible alternatives.[47] This accords with our fallibilistic understanding of reflective equilibrium.

5. Credibility and Prejudice

Copp's objection to the use of considered judgments by reflective equilibrium theorists is not the only one that has been raised. I will proceed in this section to consider two other important criticisms.

R. B. Brandt has claimed that considered moral judgments lack *initial credibility:*

> the theory [of reflective equilibrium] claims that a more coherent system of beliefs is better justified than a less coherent one, but there is no reason to think this claim is true unless some of the beliefs are initially credible – and not merely initially believed – for some reason other than their coherence, say, because they state facts of observation. … The fact that a person has a firm normative conviction gives that belief a status no better than fiction. Is one coherent set of fictions supposed to be better than another?[48]

In using the word 'fiction' I take it that Brandt does not mean that moral beliefs are by definition falsehoods. That claim would be far too strong. He means that moral beliefs are of equivalent status to invented propositions: they may be true or they may be false. If this is allowed then the complaint can be interpreted in either of two ways: first, that someone's belief in a moral proposition is *no* evidence for or against its credibility; alternatively, that someone's belief in a moral proposition is *insufficient* evidence of its credibility. Allow us to begin with the first proposition. I am unsure if it is true even of moral propositions in general, given that moral subjectivism could be true.[49] The suggestion is implausible when applied to considered moral judgments. Rawls, let us recall, describes these 'as those judgments in which our moral capacities are most likely to be displayed without distortion'. This points to the possibility that the clearing up of distortions in our judgments takes us closer to moral truth. If some kind of 'considered subjectivism' is true then this possibility may be realized. Given the fact that we cannot be sure that such a metaethical view is mistaken this makes the judgments more likely to be true than otherwise evidentially identical unbelieved alternatives (assuming, reasonably, that clearing distortions does not actually tend to take us *away* from the truth). These considerations also apply to the moral judgments held in reflective equilibrium, for they are also cleared of distortions.[50] But these judgments are more credible still as there is the further possibility that the reflective process takes us towards moral truth. This may be the case if some kind of realism whose truths we are only able to comprehend by means of philosophical reflection is true. There are, then, good metaethical reasons for believing that someone's fully reflective belief in a moral proposition gives that proposition some credibility.[51] Of course, that a proposition is believed does not show a proposition to be true; but it is evidence of its truth, and evidence is what justification is about.[52]

But is this evidence of truth sufficient? This is what the second reading of Brandt denies. The evidence provided by belief is of course weak and can be overridden by other evidence where it is stronger, as it almost always will be; in real cases it will rarely if ever be the only reason that an individual believes something. But that evidence is nevertheless sufficient to put believed judgments on a better footing than unbelieved propositions. Where there is no non-belief based evidence in favour of any propositions this may be decisive. As Brandt has offered no evidence of this kind at all – there are no ‘facts of observation’ that can show us what is right or wrong – we do not need a stronger case in order to give reflective beliefs the kind of role they have in reflective equilibria. Furthermore, in so far restricting ourselves to arguments about the evidence of *truth* provided by belief, we have granted more ground to Brandt than is required. In order for a belief to be credible it need, by definition, only be *worthy of belief*. Producing evidence that a belief is true is one way to increase its worthiness, but other considerations are at least arguably relevant. For instance, that a moral judgment comports with an individual’s reflectively favoured background theories might be thought to increase its worthiness. Arguments about the link between an individual’s beliefs and their identity might be invoked here.[53] In sum, then, it is neither the case that considered moral judgments, either prior to or in equilibria, have no credibility, nor that they are insufficiently credible.

The critical standard embodied by wide and radical reflective equilibrium should be sufficient to side-step the well-worn criticism that judgment-based moral philosophy ‘merely serves up the “conflicting ideas and feelings” that happen at any given time to predominate on the subject’.[54] Nevertheless, some will still think that they have gone too far down the path of intuitionism. Of reflective equilibrium in general R. M. Hare claims that it is merely a process of ‘gauging’ intuitions and that ‘[t]he equilibrium they have reached is one between forces which might have been generated by prejudice, and no amount of reflection can make that a solid basis for morality'.[55] Peter Singer lists the questionable sources of our intuitions with relish: 'all the particular moral judgments we intuitively make are likely to derive from discarded religious systems, from warped views of sex and bodily functions, or from customs necessary for the survival of the group in social and economic circumstances that now lie in the distant past'.[56]

The allegation is that judgments, however considered they may be, reflect *prejudice*. We evidently cannot overcome this charge simply by saying that any prejudices will be removed from considered moral judgments as they involve logical or empirical errors. Some will but many will not. Some of the damage of the criticism could be forestalled by limiting ourselves to claiming that the principles reached in reflective equilibrium are only principles for the society in which they were created.[57] Happily this is unnecessary as wide and radical reflective equilibrium systematically challenges prejudice. The individual is required to analyze each of their moral beliefs and consider how they relate to all available moral principles and arguments. She is required to look around and see the views of other people, both in her society and outside it, and try to explain any differences. She must undergo any experiences that may offset biases in her formative influences. She will also keep in mind the implications of various accepted background theories, including theories that correlate relationships between an individual's pre-philosophical beliefs and the social environment they are brought up in. (I assume that philosophical argument and empirical evidence can demonstrate the credibility of certain theories of this kind. If they cannot then the critics' case is undermined.) Consequently, if I initially hold a moral judgment merely because it reflects outmoded views on religion or sex I am likely to realize that and consequently reject it.

Of course it is possible that in some cases plausible explanations of prejudicial views may be available, and that the prejudices are therefore retained in a rationalized form. But the method of reflective equilibrium remains attractive as it has mechanisms in place to attempt to minimize the occurrence of such instances. Where prejudice runs so deep that it will be held on to after the fullest reflection there is very little that can be done to remove them.[58] The possibility of enduring prejudice does not count against the method; it simply prevents it, together with all other methods of justification in ethics, from legitimately claiming that it enables individuals to make moral judgments from outside of their moral tradition.

6. Fair Procedure

Some critics accept judgments as a starting point for moral enquiry, but maintain that the method of reflective equilibrium is a poor way of using them. A common claim is that the procedural requirements of the reflective process such as coherence are insufficiently discriminating. Haslett asks us to suppose we are unable to decide between two mutually contradictory principles: 'All coherence considerations enable us to decide is that the one *or* the other must be chosen, they do not enable us to decide, definitively, *which* one'.[59] Consequently such decisions must be made in an arbitrary fashion. Likewise, Joseph Raz writes that reflective equilibrium takes us 'back to the usual philosophical argument about the merits and demerits of various methods of argument and of various theories. The method of reflective equilibrium is then not a method in moral philosophy at all. It simply advocates that our judgment be informed … and consistent.'[60]

The question ultimately turns on the availability of alternative methods of justification. Haslett is content to limit himself to the specification of ‘one which, like reflective equilibrium methodology, requires coherence and consistency with all known facts but, unlike reflective equilibrium methodology, does *not* tolerate circularity and underdetermined adjustment decisions’.[61] However, these two claimed dissimilarities are deeply problematic. The claim of circularity is reliant on the argument against the independence constraint addressed above. The wish for some definitive non-judgment based way of choosing one moral conception over another is hopeless without knock-down, judgment-trumping reasoning – the existence of which I have denied. Raz accepts reflective equilibrium’s methodological ‘proposition that the less likely one is to abandon a moral belief just because one has acquired more information, the more trustworthy that belief is’ and its limitless modification of judgments, but objects that ‘it fails in a basic requirement of a method of moral argument, that is the ability to guide the agent’s choice of moral views’.[62] Like Haslett, Raz expects too much of a method of moral justification. No such method can hope to provide a judgment-independent means of arbitration; all that we can hope for is a reasonable way of coping with moral judgments.

As I have tried to show, the method of wide and radical reflective equilibrium provides a *fair procedure* for handling considered judgments. The rejection of the confidence constraint ensures that all judgments are accounted for; the defence of the independence constraint ensures that no judgments are double-counted; the consideration of all arguments for every principle ensures that the best case is presented for each; the attention paid to others’ views ensures that we consider possible causes of us holding the views we do; and the requirement to offset formative biases ensures that, insofar as is possible and reasonable, we make our judgments from a balanced standpoint. Scanlon's comment that the method 'prescribes, so to speak, a level playing field of intuitive justification on which principles and judgments of all levels of generality must compete for our allegiance' is, then, especially pregnant when applied to the variant defended here.[63] Once the most obvious distortions (caused by emotional distractions and factual and logical errors) are cleared up, all that is prohibited is the identification of some set of judgments or principles that do not require support from considered judgments in order for them to figure in our conclusions. In an ideal philosophical world we would have a more discriminating method of moral justification. But I hope to have shown that, given the actual philosophical world, with all its uncertainties, an appropriately modified variant of the method of reflective equilibrium is a fair and plausible way for us to choose our moral principles.

It is a separate question whether the method is practicable. Rawls suggested that something fairly close to my position would be 'the philosophical ideal' but that 'it is doubtful that one can ever reach this state'.[64] Time pressures may indeed make it necessary to apply the confidence constraint, to consider only two or three rival theories, to forego most of the available transformative experiences, and so on. But the demands of practicality cannot completely override the demands of philosophy. Pure moral epistemology is itself not without value. And if we must make compromises, it is better to understand just what is being compromised. When taking shortcuts, it helps to know the lay of the land.

Notes

[*] This is a pre-publication version of Carl Knight, 'The Method of Reflective Equilibrium: Wide, Radical, Fallible, Plausible', *Philosophical Papers,* 35 (2006), 205-229. The published version is here: https://www.tandfonline.com/doi/abs/10.1080/05568640609485180 For their helpful comments and suggestions, I thank Matt Matravers, Hillel Steiner and two anonymous referees. Research for this article was supported by the Arts and Humanities Research Council.

[1] I will not argue for this point. For brief comments on Brandt's (1979) attempt at such a derivation see note 58 below. Of the many contemporary attempts of this general kind, Hare's (1981) is probably the best-known. Important criticisms of it are presented by McDermott (1983), Pettit (1987) and Nagel (1988). Contractarian (Gauthier, 1986) and communitarian (MacIntyre, 1985) attempts have also had some influence; I say something about the latter in section 3.

[2] cf. Daniels, 1996; Nagel, 1991, 7-8.

[3] Rawls, 1999, sec. 9; 1975.

[4] Rawls, 1999, 42.

[5] Rawls, 1999, 42.

[6] Rawls, 1999, 42.

[7] The judgment must not be, in Williams' phrase, 'conflicting' – i.e. logically consistent but inconsistent with empirical facts (1988, 41-2).

[8] Rawls, 1999, 42.

[9] Scanlon, 2003, 144.

[10] Raz, 1982, 123.

[11] Miller, 1999, 56.

[12] Rawls, 1999, 42-3.

[13] See Brink, 1989, 130; Kagan, 1989, 12-4; Schroeter, 2004, 118-121.

[14] Daniels, 1996, 67-9.

[15] Rawls, 1999, 49; 1975, 8.

[16] Daniels, 1996, 23.

[17] DePaul 1988, 84.

[18] Haslett, 1991, 138.

[19] Daniels, 1996, 23.

[20] Daniels, 1996, 23-4; Rawls, 1999.

[21] Haslett, 1987, 139.

[22] See Rawls, 1999, 46, sec. 27-30.

[23] See Raz, 1982, 133; cf. Schroeter, 2004.

[24] Rawls, 1999, 44.

[25] Miller, 1999, 55.

[26] Rawls, 1999, 44.

[27] Walzer, 1985, 29; cf. MacIntyre, 1985.

[28] Daniels, 1996, 109-10.

[29] DePaul, 1987, 466.

[30] DePaul, 1987, 463.

[31] DePaul, 1987, 470.

[32] I am less sure that DePaul is correct in maintaining that the conservative interpretation, or something like it, is the dominant conception of reflective equilibrium. Rawls certainly encouraged the revision of initial judgments where this was suggested by philosophical reflection (see, e.g., Rawls 1999, 43). He does not, however, consider the role of non-philosophical sources. Neither of DePaul 1987's interpretations can therefore be unproblematically attributed to him. Rawls' stance actually seems to fit the different, less radical definition of radical reflective equilibrium given in DePaul's book *Balance and Refinement* (1993). In this work the more radical interpretation that gives a role to literature, film, art and music is referred to as 'the method of balance and refinement'.

[33] DePaul, 1987, 471.

[34] The only class of such activities to be avoided are those that are, in a special sense, corrupting (see DePaul, 1993, ch. 4).

[35] Brink, 1989, 105

[36] See Daniels, 1996, 60-1.

[37] DePaul, 1986; Ebertz, 1993.

[38] Daniels, 1996, 27, original emphasis.

[39] See Brink, 1989, 134; Ross, 1930, 41-2.

[40] Harman, 2003: 415.

[41] Brink, 1989, 134.

[42] In this article I do not consider multi-agent and multi-equilibria cases.

[43] Ebertz, 1993, 204.

[44] See Brink, 1989, 132-3.

[45] Copp, 1986, 165.

[46] Copp (1986, 161-2, 165) does himself no favors here by mentioning charisma and 'pressure to conform'. Insofar as such factors affect our judgment, that judgment is not genuinely reflective.

[47] See Brink, 1989, 128-30.

[48] Brandt, 1979, 20.

[49] By moral subjectivism I mean the radical view that moral propositions derive all (or much) of their truth-value from being believed. As we cannot be certain that this view is incorrect, someone's belief in a moral proposition therefore increases the likelihood of its being true above that of some unbelieved proposition. (I assume that if subjectivism is mistaken, someone's belief in a moral proposition does not *decrease* its likelihood of being true.)

[50] Brandt's complaint seems to be directed at considered moral judgments before the reflective process begins (see Brink, 1989, 135). This is perhaps because Brandt has some kind of conservative reflective equilibrium in his sights. But the credibility complaint could be made, *mutatis mutandis,* against radical reflective equilibrium, hence my response.

[51] In the text I have not exhausted such reasons. For instance, merely clearing distortions may tend to take us towards truth if realism is correct, or the reflective process may take us towards truth if 'reflective subjectivism' is true.

[52] Brink, 1989, 125-6, 140.

[53] cf. DePaul, 1993.

[54] Callinicos, 2000, 64.

[55] Hare, 1981, 12; cf. 1975.

[56] Singer, 1974, 516, cf. Brandt, 1979, 21-2.

[57] See Williams, 1985, 102, original emphasis; cf. Rawls, 1993. But even were we satisfied with limiting ourselves in this way the critics' charge is not fully met, for prejudices about matters internal to the society would still go unchallenged.

[58] In this respect the method has the same inevitable limitations as Brandt's 'cognitive psychotherapy', which attempts to make desires rational by confronting them 'with relevant information, by repeatedly presenting it, in an ideally vivid way, and at an appropriate time' (Brandt, 1979, 113). The individual remains situated, having been shaped by social institutions, which have themselves been shaped by a moral code (Daniels, 1996, 93).

[59] Haslett, 1987, 141, original emphasis.

[60] Raz, 1982, 112-3.

[61] Haslett, 1987, 143, original emphasis.

[62] Raz, 1982, 133.

[63] Scanlon, 2003, 151.

[64] Rawls, 1999, 43; also 1993, 97.

References

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

Brink, David O. 1989. *Moral Realism and the Foundations of Ethics.* Cambridge: Cambridge University Press.

Callinicos, Alex. 2000. *Equality.* Cambridge: Polity.

Copp, David. 1985. 'Considered Judgments and Moral Justification: Conservatism in Moral Theory.' In *Morality, Reason and Truth: New Essays on the Foundations of Ethics,* ed. David Copp and David Zimmerman. Totowa, Nj.: Rowman & Allanheld.

Daniels, Norman. 1996. *Justice and Justification: Reflective Equilibrium in Theory and Practice.* Cambridge: Cambridge University Press.

DePaul, Michael R. 1986. 'Reflective Equilibrium and Foundationalism.' *American Philosophical Quarterly* 23: 59-69.

---. 1987. 'Two Conceptions of Coherence Methods in Ethics.' *Mind* 96: 463-81.

---. 1988. 'The Problem of the Criterion and Coherence Methods in Ethics.' *Canadian Journal of Philosophy* 18: 67-86.

---. 1993. *Balance and Refinement: Beyond Coherence Methods of Moral Enquiry.* London: Routledge.

Ebertz, Roger P. 1993. 'Is Reflective Equilibrium a Coherentist Model?' *Canadian Journal of Philosophy* 23: 193-214.

Gauthier, David. 1986. *Morals by Agreement.* Oxford: Oxford University Press.

Hare, R. M. 1975. 'Rawls' Theory of Justice.' In *Reading Rawls: Critical Studies on Rawls'* A Theory of Justice*,* ed. N. Daniels. Oxford: Blackwell.

---. 1981. *Moral Thinking: Its Levels, Method, and Point.* Oxford: Oxford University Press.

Harman, Gilbert. 2003. 'Three Trends in Moral and Political Philosophy.' *The Journal of Value Inquiry* 37: 415-25.

Haslett, D. W. 1987. 'What is Wrong with Reflective Equilibria?' In *Equality and Liberty: Analyzing Rawls and Nozick,* ed. J. Angelo Corlett. Basingstoke: MacMillan.

Kagan, S. 1989. *The Limits of Morality.* Oxford: Oxford University Press.

MacIntyre, Alasdair C. 1985. *After Virtue: A Study in Moral Theory,* second edition. London: Duckworth.

McDermott, Michael. 1983. 'Hare's Argument for Utilitarianism.' *The Philosophical Quarterly* 33: 386-91.

Miller, David. 1999. *Principles of Social Justice.* Cambridge, MA: Harvard University Press.

Nagel, Thomas. 1988. 'The Foundations of Impartiality.' In *Hare and Critics: Essays on Moral Thinking,* ed. Douglas Seanor and N. Fotion. Oxford: Oxford University Press.

---. 1991. *Equality and Partiality.* Oxford: Oxford University Press.

Pettit, Philip. 1987. 'Universalizability Without Utilitarianism.' *Mind* 96: 74-82.

Rawls, John. 1975. 'The Independence of Moral Theory.' *Proceedings and Addresses of the American Philosophical Association* 47: 5-22.

---. 1993. *Political Liberalism.* New York: Columbia University Press.

---. 1999. *A Theory of Justice,* revised edition. Oxford: Oxford University Press.

Raz, Joseph. 1982. 'The Claims of Reflective Equilibrium.' In *Equality and Liberty: Analyzing Rawls and Nozick,* ed. J. Angelo Corlett. Basingstoke: MacMillan.

Ross, W. D. 1930. *The Right and the Good.* Oxford: Clarendon Press.

Scanlon, Thomas. 2003. 'Rawls on Justification.' In *The Cambridge Companion to Rawls,* ed. Samuel Freeman. Cambridge: Cambridge University Press.

Schroeter, François. 2004. 'Reflective Equilibrium and Antitheory.' *NOÛS* 38: 110-34.

Singer, Peter. 1974. 'Sidgwick and Reflective Equilibrium.' *The Monist* 58: 490-517.

Walzer, Michael. 1985. *Spheres of Justice: A Defense of Pluralism and Equality.* Oxford: Blackwell.

Williams, Bernard. 1985. *Ethics and the Limits of Philosophy.* London: Fontana.

---. 1988. 'Ethical Consistency.' In *Essays in Moral Realism,* ed. Geoffrey Sayre-McCord. Ithaca, NY: Cornell University Press.