Are reflective equilibrium and the original position consistent? The historical bias problem

Author: Terence Rajivan Edward

Abstract. In this paper, I present a problem for regarding the reflective equilibrium and original position methods as consistent. I do not prove that there is an inconsistency, but there is a puzzle of how the two methods can be made consistent. The concern about inconsistency is because the former method allows for a kind of historical bias, as noted by T.H. Irwin, whereas the latter method seeks to guard against historical bias.

Draft version: Version 3 (October 31st 2022, table edit).

John Rawls famously recommends at least two methods when working out what the government in a liberal society should do, namely reflectively equilibrium and the original position. These are impressive contributions to philosophy and both methods seem indispensable, but there is a question of whether the two methods are consistent. I shall begin by outlining the reflective equilibrium method before discussing an objection raised by T.H. Irwin, which will reveal a certain commitment that one must defend in order to defend Rawls's version of the method. I shall then outline the original position method and draw attention to a commitment that is needed in order to defend his version of that method. But whatever argument supports the former commitment would seem to be inconsistent with whatever argument supports the latter commitment.

Here is a rough description of the reflective equilibrium method. It involves a person, or set of persons, beginning with moral judgments about specific situations, for example the government should not burn Ned's book for no reason. The aim of the method is to systematize these judgments, that is to say, to identify a general principle or a few general principles from which one can infer all of these judgments. If one proposes a principle, such as that the government should try to produce as much happiness as possible, and from this principle one can only infer half of these judgments, then one has to change the proposal. But if one's proposed principle enables one to infer almost all of these judgments, then there is the option of abandoning either the principle or those judgments which do not fit.[1] The judgments entered into the procedure do not have an unquestionable status in contrast to proposed principles. Once one has achieved a total fit, whether by revising proposals or revising judgments, then one has achieved the end aimed at by the method – one has achieved equilibrium (1999: 18). General principles count as justified in this condition, according to the method.

T.H. Irwin raises an important objection. If one's proposed general principles entail most of one's specific judgments, there is the option of abandoning judgments which do not fit, and thereby adapting judgments to fit theory; but what if rival principles, proposed later, fit with the earlier set of judgments, which include the abandoned judgments? By the standards of reflective equilibrium, the rival principles look worse but merely because they were proposed later – they are now being evaluated by how they fit with a set of specific judgments which have been adapted for the principles proposed earlier. To illustrate this point – the table below may be clearer than plain prose, by the way – let us suppose that a person starts with only J1, J2, J3, and J4 as their specific judgments and they propose P1 and P2 as their general principles. But J4 does not fit with these principles, so the person abandons J4 and they replace it with J5. Now there is equilibrium, but someone comes along and introduces the person to a rival set of general principles, composed of P1 and P3. This rival set fits with J1, J2, and J3, but it does not fit with

[1] For the sake of simplicity, I shall here assume that specific judgments are either validly inferable from a proposed set of general principles or inconsistent with these principles.

J5, so they reject it in favour of their current set of principles, because this current set fits with all their specific judgments. But the rival set does fit with J4. So allowing for revision of judgments enables one set of general principles to score more highly than another, using the reflective equilibrium method, merely because that set was conceived first and then one's judgments were adapted to achieve a fit with those general principles. If both sets of general principles were conceived before the revision process, the reflective equilibrium method would not lead to this justificatory rating. (The judgments about specific situations are supposed to be analogous to data used to justify scientific theories, but we don't want a method of justification which allows one scientific theory to score better merely because it was conceived earlier.)

Time	**Reflective equilibrium situation, for an individual**
July 2022	Judgments about specific situations: J1, J2, J3, and J4. No proposed general principles – no reflective equilibrium achieved.
August 2022	Judgments about specific situations: J1, J2, J3, and J4. Proposed general principles: P1 and P2. Principles entail: J1, J2 and J3 only – still no reflective equilibrium.
September 2022	J4 abandoned for J5 to achieve reflective equilibrium. Judgments about specific situations: J1, J2, J3, and J5. Proposed general principles: P1 and P2. Principles entail: J1, J2, J3, and J5 – reflective equilibrium.
October 2022	Judgments about specific situations: J1, J2, J3, and J5. General principles: P1 and P2, reflective equilibrium achieved. Rival system of general principles conceived: P1 and P3. Rival system entails: J1, J2, J3, but not J5. Rival system fails to achieve reflective equilibrium. Rival system also entails J4. It would have achieved reflective equilibrium if proposed in July or August 2022, thereby preventing abandonment of J4 in September 2022.

Irwin's solution is that what we need is a set of judgments that we cannot revise (2009: 901), but then the reflective equilibrium method would lose its distinctive quality of not giving absolute priority to either principles or judgments (1999: 18). It would change from what epistemologists call a coherentist procedure, in which justification is achieved by making various "elements" cohere with no kind of element accorded priority, to a foundationalist procedure: a procedure in which one kind of element is the foundation – in this case the unrevisable judgments. Knowledge is achieved by examining whether one's principles fit with this foundation. To defend Rawls's coherentist version of reflective equilibrium, it seems that one needs to defend the following commitment:

> (*Historical bias commitment*) If justification by reflective equilibrium is achieved between a set of general principles and a set of specific judgments by revising the judgments, and if a rival set of principles conceived later is less justified given the revised set of judgments, but not with the unrevised set, this is not a problem.

This is where a question arises of how to achieve consistency with the original position method.

The original position method involves imagining a set of self-interested individuals choosing from a menu of options, each option being a set of general principles. They are self-interested, so they choose to implement the principles that are most in their interests. But these self-interested individuals lack certain information about themselves. For example, they do not know their sex, their wealth, talents, or aspirations. The idea is that we should implement fair principles and by depriving them of this knowledge, the result is that they choose fair ones, because no individual chooses principles which favour features of themselves which are not shared. One would not choose that only males get to govern, or only females do, because one does not know one's sex or talents or aspirations in life. For all one knows, one might want to go into government and a principle that only allows one sex to govern might prevent that. Now the principles chosen are supposed to be implemented over generations, so the individuals involved also do not know details about which period they are living in. A full defence of the original position method is therefore going to involve a defence of the following commitment:

> *(Anti-historical bias commitment)* If individuals in the original position know details about which historical period they are living in, this would prevent the fairness that this method is trying to achieve.

But would not any argument which is good enough to support this commitment be inconsistent with any argument which is good enough to support the historical bias commitment identified earlier? There is a problem of how you reconcile the two methods, one of which requires accepting historical bias and the other of which takes measures to prevent this.

References

Irwin, T.H. 2009. *The Development of Ethics: A Historical and Critical Study*. *Volume III: From Kant to Rawls*. Oxford: Oxford University Press.

Rawls, J. 1999 (revised edition). *A Theory of Justice.* Cambridge, Massachusetts: Belknap Press.