Phil. 262: Intuitions and Philosophical Methodology
Nov. 20, 2007

What Is Reflective Equilibrium?

I. Three Reflective Equilibriums

"Reflective equilibrium" can refer to three different things:

- A *state* that one's beliefs might be in.
- A *method* that one might use for doing philosophy in some area.
- A *theory* that states what it takes for one's beliefs about some philosophical domain to be *justified*.

Today we will be primarily concerned with the first and second of these.

II. The Method of Narrow Reflective Equilibrium

A recipe for carrying out the *method of narrow reflective equilibrium* in the field of normative ethics:

i. Start with your *initial moral judgments*.

ii. Filter these to get your *considered moral judgments*.

iii. Propose a set of *moral principles* that explicate your considered moral judgments.

iv. If there is a conflict between
   a. your set of considered moral judgments, and
   b. your set of moral principles,

   then revise one of the conflicting elements.

v. Repeat stage (iv) above until a state of equilibrium is reached.

Some comments:

- As formulated, the method is to be used by a *single individual*.
- Presumably the judgments in question are judg*ed*s, not judg*ing*s.
- The *initial moral judgments* are at all levels of generality, and include particular-case judgments with their supporting reasons. Examples:

  "What he did was wrong because it hurt her."
  "One ought to keep one's promises."
  "Any two actions sharing all of their non-moral properties must also share their moral properties."

- The *considered moral judgments* are filtered to exclude the following sorts of initial moral judgments:

  1. Judgments made when one is not aware of the relevant (non-moral) facts about the issue at question.
  2. Judgments made when one is upset, frightened, or otherwise not able to concentrate.
  3. Judgments made when one stands to gain or lose on the basis of the answer given.
  4. Judgments made with hesitation, or with little confidence.
  5. Judgments that are not stable over time.

- The *moral principles* proposed should be such that "when conjoined to our beliefs and knowledge of the circumstances, would lead us to make these judgments with their supporting reasons were we to apply these principles conscientiously and intelligently" (Rawls, *A Theory of Justice*, p. 46).

- As formulated, it is left open how one is to go about resolving any conflicts found in stage (iv).
- The considered judgments are not fixed points: "there are no judgments on any level of generality that are in principle immune to revision" (Rawls, "The Independence of Moral Theory," p. 8).
- We might require that the final equilibrium state be *coherent*, where this entails more than *consistency*; for example, we might require the moral principles to possess explanatory virtues such as *simplicity*.

> "Moral philosophy is Socratic: we may want to change our present considered judgments once their regulative principles are brought to light. And we may want to do this even though these principles are a perfect fit" (Rawls, *A Theory of Justice*, p. 49).

- No assumption is made that we will ever, in fact, reach a final state of equilibrium.
- Also, no assumption is made that there is a *unique* equilibrium point.

III. The Method of Wide Reflective Equilibrium (Rawls' Version)

Rawls on the distinction between *narrow* and *wide reflective equilibrium*:

> "There are, however, several interpretations of reflective equilibrium. For the notion varies depending on whether one is to be presented with only those descriptions which more or less match one's existing judgments except for minor discrepancies, or whether one is to be presented with *all possible descriptions* to which one might plausibly conform one's judgments together with *all relevant philosophical arguments for them*" (Rawls, *A Theory of Justice*, p. 49, emphasis added).

> ". . . we are interested in what [moral] conceptions people would affirm when they have achieved *wide* not just *narrow* reflective equilibrium, an equilibrium that satisfies certain conditions of rationality. That is, . . . we investigate what principles people would acknowledge and accept the consequences of when they have had an opportunity to consider other plausible conceptions and to assess their supporting grounds. Taking this process to the limit, one seeks the conception, or plurality of conceptions, that would survive *the rational consideration of all feasible conceptions and all reasonable arguments for them*" (Rawls, "The Independence of Moral Theory," p. 8, emphasis added).

A reason for preferring *wide* over *narrow* reflective equilibrium: it helps combat the *charge of conservativism*.

> When practicing wide reflective equilibrium, one's moral views can undergo quite radical shifts. Moreover, wide reflective equilibrium seems better suited than narrow reflective equilibrium at correcting for initial moral judgments that are the product of *bias*, *historical accident*, or *ideology*.

This advantage may come at a price, namely that of opening the method to the *charge of emptiness*:

> "But once the method has been broadened in the way just described, . . . it seems to become empty as a methodological doctrine. It becomes simply the truism that we should decide what views about justice to adopt by considering the philosophical arguments for all possible views and assessing them on their merits" (Scanlon, "Rawls on Justification," p. 151).

IV. The Method of Wide Reflective Equilibrium (Daniels' Version)

Daniels construes the philosophical arguments for the alternative moral conceptions to be inferences from a relevant *body of background theories*, such as a theory of personal identity, a theory of the role of morality in society, a theory of meaning, and so on.

Thus Daniels interprets the *method of wide reflective equilibrium* as an attempt to produce coherence in an ordered triple of sets of beliefs held by a particular person, namely:

a. a set of considered moral judgments,
b. a set of moral principles, and
c. a set of relevant background theories.

Three reasons to be suspicious of whether Daniels' interpretation of Rawls is accurate:

- The whole point of "The Independence of Moral Theory" is that moral theory is (largely) independent from epistemology, the theory of meaning, and the problem of personal identity.
- On Daniels' interpretation, *non-moral* background theories are allowed to change during the process of achieving *moral* reflective equilibrium; Rawls never explicitly mentions this.
- We don't seem to be directly considering alternative moral conceptions on Daniels' version.

Nevertheless, Daniels' interpretation of wide reflective equilibrium has become the standard one, and anyway it represents a methodological view that is worthy of consideration in its own right.

Several reasons why the set of background theories will be much vaster than Daniels suggests it is:

- The set of directly relevant theories is larger than advertised (for example, it no doubt includes large swathes of metaphysics, epistemology, the theory of action, the philosophy of mind, etc.).
- Not only must we include all theories directly relevant to moral theory, but also all theories directly relevant to the theories directly relevant to moral theory (for amending a directly relevant theory could have consequences elsewhere that should be taken into account).
- Also, we must include all theories directly relevant to the theories directly relevant to the theories directly relevant to moral theory, and so on.

Indeed, the case can be made that Daniels has smuggled *all the rest of philosophy* into the reflective mix.

If the set of background theories includes the central tenets of reflective equilibrium (either as a theory of justification or as a method of philosophy), this gives rise to some intriguing possibilities.

> What happens if we decide to revise some of the basic tenets of reflective equilibrium during the course of practicing the method of reflective equilibrium?
>
> For example, we might decide to revise our judgment, "No judgment is immune to revision." But then it seems that it was all along possible for a judgment to be immune to revision: we just had to revise our judgment "No judgment is immune to revision" first.

V. Daniels' Independence Constraint

In order for one's ordered triple <(a), (b), (c)> to be in a state of wide reflective equilibrium, Daniels thinks that it's not enough that it be fully coherent; in addition, it must satisfy the following constraint:

> *the independence constraint*: "Some interesting, nontrivial portions of the set of considered moral judgments that constrains the background theories and of the set that constrains the moral principles should be disjoint" (Daniels, "Wide Reflective Equilibrium...," p. 259).

The idea: let (a)* be the subset of (a) that constrains (b), and let (a)** be the subset of (a) that constrains (c). Then (a)* and (a)** should be to some significant degree disjoint.

Daniels thinks that once the independence constraint is in place, wide reflective equilibrium avoids the possibility that the moral principles in a state of equilibrium are just "accidental generalizations" of the moral facts, analogous to accidental generalizations that we want to distinguish from real scientific laws.

I find it difficult to understand what Daniels means by one set "constraining" another:

> Since the moral principles are supposed to generate all of the considered moral judgments, there seems to be a sense in which *all of (a)* constrains (b) when one is in a state of equilibrium.
>
> Or maybe a considered moral judgment counts as constraining (b)/(c) if at some time in the past, one revised (b)/(c) because an element in it conflicted with that considered moral judgment.

VI. Doubts about the Filtering Process

One might have worries about the epistemic defensibility of the initial filtering process:

> Against ruling out judgments of type 2: judgments made out of moral outrage might more accurately track egregious wrongdoings than judgments made in a calm emotional state.
>
> Against ruling out judgments of type 3: doing so seems to stack the deck against ethical egoism.
>
> Against ruling out judgments of type 4 and 5: "Ideally, we would want to find principles that explain these judgments as well" (Scanlon, "Rawls on Justification," p. 144).

Indeed, the inclusion of this initial filtering process might seem at odds with the general spirit of reflective equilibrium, since the epistemic principles that underwrite the different sorts of filtering seem to have been elevated to the status of unimpeachable principles immune to revision.

A suggestion: Instead of filtering the initial moral judgment, include the epistemic principles that underwrote the filtering process in the set (c) of background theories. Then these can be given up or revised as seems appropriate.

VII. How Do We Decide What Changes to Make?

Suppose we identify a conflict between one of our considered moral judgments and one of our moral principles. Rawls writes, "In this case we have a choice": either we revise the considered judgment, or we revise the principle (*A Theory of Justice*, p. 20). However, how do we go about making this choice?

Three salient options present themselves:

1. Arbitrarily choose which one to change.
2. Make the change that *seems* most plausible or reasonable.
3. Make the change that *as a matter of fact* is most plausible or reasonable.

Before addressing this issue, let's address a related one:

> When does a choice need to be made? Is it enough that, as a matter of fact, there *is* a conflict between two (or more) elements of sets (a), (b), and (c)? Or does one have to *believe* that there is a conflict between two (or more) elements of sets (a), (b), and (c)? Against the latter option:
>
> - Then reflective equilibrium would seem to be too easy to achieve: one could attain equilibrium merely by being logically incompetent or intellectually lazy.
>
> - It leads to a Lewis Carroll-style regress. Suppose one believes moral principle *m* and considered judgment *j*, and suppose the following is true:
>
> (*) *m* and *j* are jointly inconsistent.
>
> The claim is that the truth of (*) is not enough to rationally mandate that one revise either *m* or *j*; in addition, one must believe (*). However, it seems that one's belief in *m*, *j*, and (*) only mandates a revision because the following is true:
>
> (**) *m*, *j*, and (*) are jointly inconsistent.
>
> But if the brute truth of (**) is enough to mandate a revision, why isn't the brute truth of (*) enough? Thus either we should require that one also believe (**) (and so on, ad infinitum), or else we should rethink the proposal being considered.

Thus reflective equilibrium, on its most plausible interpretation, seems to allow at least some non-psychological facts to play a role in how one should update one's beliefs.

But if we allow external norms that are potentially out of one's purview to determine when one must make a change, why not allow external norms to determine how one should go about making that change?