Rawlsian Reflective Equilibrium

Abstract: This paper proposes a Rawlsian conception of moral justification as a social activity. Through a close reading, Rawls' view of ethical justification is shown to be significantly more *dialogical* and *deliberative* than is commonly appreciated. The result is a view that emphasizes the social nature of ethical justification and identifies information sharing between persons as the crux of justification in metaethics, in contrast to normative ethics. I call it *Rawlsian reflective equilibrium* to distinguish it from other varieties.

Keywords: reflective equilibrium, moral justification, deliberation, Rawls

Word Count: 3,327

1. Reflective Equilibrium and its Discontents

Reflective equilibrium (RE) is a popular[1] procedural account of moral and ethical justification.[2] RE requires that an agent (i) begins in a state with certain initial moral judgments; (ii) proceeds to prune these based on her logico-empirical standards, resulting in a set of considered moral judgments; (iii) adduces moral principles that explicate those judgments; (iv) brings the set of judgments and principles into coherence by selecting among the members of the united set, according to the criterion of maximal consistency; and, (v) brings this latter set into maximal coherence with other relevant background theories (following DePaul 1993, 16-23).

Despite its popularity, RE has received ample criticism. The general issue lies with the method's apparent circularity: RE appears to rest entirely upon the judgments and principles adopted by the agent instantiating it, so whatever judgments, principles, or intuitions she begins with biases the composition of the "justified" set at equilibrium (*e.g.*, Haslett 1987, 307). In response, proponents distinguish a type of RE that is admittedly circular, though in a putatively unobjectionable sense. Yet critics remain unsatisfied. They argue this variety either demands too much from cognitive agents and hence is impracticable, or it again falls to the charge of problematic circularity (DePaul 1993).

[1] For example, RE figures centrally in canonical works in bioethics (*e.g.*, Daniels 1979, 1996; Beauchamp and Childress 2009), so much so that Arras claims, "in the world of bioethics, the air is abuzz with reflective equilibrium" (2007, 46).

[2] RE may be distinguished from a related epistemological thesis (cf. Stitch 1988), which will not concern us.

This paper proposes a novel interpretation of Rawls' account of reflective equilibrium in response to these criticisms. By showing Rawls' view to be far more dialogical and deliberative than most appreciate its analysis makes important progress in our understanding of RE, providing textual support for a new understanding of Rawlsian ethical justification and aligning it with growing work on the social nature of rationality and cognition.

2. Varieties of Reflective Equilibrium

In addition to Rawls, Norman Daniels is also credited with developing reflective equilibrium as a method of moral justification. It is important to distinguish his account from Rawls' because they differ in important ways.

According to Daniels, reflective equilibrium is a method of justification that attempts "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" (1979, 258). Though this method might appear to begin with {a}, move to {b}, and conclude with {c}, this is only apparent: the individual performing reflective equilibrium may begin with either set and move in between them in any order she likes, just so long as producing coherence is the activity guiding her reasoning (*ibid.*, 259 fn. 5).

Daniels recognizes that as formulated above RE is open to the charge of circularity. Depending upon what information must be sampled in order to have adequately "produced coherence" in {a}, {b}, and {c}, the method may merely entail making ones own idiosyncratic views coherent. Consequently, Daniels distinguishes the mere coherence of one's considered views (narrow, or NRE) from *wide reflective equilibrium* (WRE), which is designed to protect against the charge of circularity. On WRE, "we can imagine the agent working back and forth, making adjustments…[who] arrives at an equilibrium point that consists of the ordered triple (a), (b), (c)" (258ff.). To Daniels, the important feature of WRE is its appeal to "background theories," which provide independent grounds for an agent's selection of moral principles. Also, Daniels' description is decidedly personal and egocentric: "the agent" works back and forth

adjusting "his" judgments, "his" principles, and "his" background theories. And, in this way "he" arrives at equilibrium. Thus for Daniels, RE is a method employed by individual agents striving for a coherent moral theory, but only WRE carriers justificatory force because it alone sufficiently appeals to independent background theories (cf. DePaul 1993, 16-23).

3. Rawls on Reflective Equilibrium

Although Rawls did not initially distinguish between varieties of reflective equilibrium, his discussion in *A Theory of Justice* suggests to some authors that the distinction is implicit there (Haslett 1987). Yet in his later work, Rawls rarely distinguishes between varieties or deploys different senses of RE. If one adopts the view proposed here this makes sense because these senses are not as consistent with Rawls' view as others contend.

Rawls' formulation of reflective equilibrium rests on the concept of *mutual support of many considerations*. For him, the method begins with our most laudatory considered convictions. It then justifies them by exposure to criticism from as many sources as possible. In this manner, we move from our initial judgments to a set of justified principles at equilibrium. For Rawls, the semi-stable endpoint of equilibrium serves as a justification for his target, a definition of the principles of justice, which in his elaborate theory is called "the original position," referred to below as "the situation."

> In searching for the most favored description of this situation we work from both ends. We begin by describing it so that it represents generally shared and preferably weak conditions. We then see if these conditions are strong enough to yield a significant set of principles. If not, we look for further premises equally reasonable. But if so, and these principles match our considered convictions of justice, then so far well and good. But presumably there will be discrepancies. In this case we have a choice. We can either modify the account of the initial situation or we can revise our existing judgments, for even the judgments we take provisionally as fixed points are liable to revision. By going back and forth …eventually we shall find a description of the initial situation that both expresses reasonable conditions and yields principles which match our considered judgments duly pruned and adjusted. This state of affairs I refer to as *reflective equilibrium* (1999, 18; italics added).

It is important to appreciate not only how Rawls describes what RE is, but also what it is not. RE is not a procedure for making *particular* ethical decisions, nor is it for abstractly describing justified ethical decision-making (cf. Rawls 1951). For Rawls, reflective equilibrium is an ideal procedure for justifying an abstract conception of justice, which we may refer to as a 'higher-order principle' because it rests on a semi-stable, coherent system of lower-order principles and judgments that follows from iterating the RE procedure.

In his introductory account of RE, Rawls is less egocentric and logicistic than Daniels. Rawls states that "we" work from both ends, "we" describe "shared" information, and "we" see if this information suffices for yielding principles. Also, "we" go back and forth, altering information as necessary, so that "we" find a description of the situation at equilibrium. For Rawls what is important is that the right information is shared, considered, and agreed upon. He does not emphasize its logical structure, nor does he suggest that the method is something an individual reflective agent performs.

In later specifications of RE Rawls' emphasis on non-formal information sharing between agents persists. Consider *Political Liberalism*, where, for example, Rawls states "the third point of view – that of you and me – is that from which justice as fairness…is to be assessed. Here the test is that of reflective equilibrium: how well the view as a whole articulates our more firm considered convictions of political justice, at all levels of generality, after due examination, once all adjustments and revisions that seem compelling have been made" (2005, 28). If we think of a theory of political justice as merely one of any number of ethical theories or principles that could be justified by RE, then Rawls may be interpreted here as stating that forging agreement within a third, shared point of view provides the foundation of justification for them. In addition to this foundation, the procedure requires a certain means of considering what others think about the principles at issue. Elsewhere in this work, Rawls notes that the right conception of this process will be "a dialogue; indeed, an omnilogue" (*ibid.*, 383). All viewpoints ought to be included, and no viewpoint ought to be privileged before the fact.

4. Rawls on Ethical Justification

By returning to Rawls' take on the activity of ethical justification in *Theory*, we find more support for the view that Rawlsian reflective equilibrium emphasizes the sharing of reasons in a social context. It also emphasizes a natural language approach to conceptualizing the information incorporated into the RE procedure, rather than a semi-formal logical approach describing sets of beliefs brought into coherence with one another.

Rawls portrays justification as essentially *dialogical,* as opposed to logical. By his lights, empirical demonstration of truth and deduction from self-evident principles are insufficient for ethical justification; rather, what is necessary is a shared starting point, which may include merely the shared stance of engaging in argument and recognizing the authority of consensus.

> Justification is argument addressed to those who disagree with us, or to ourselves when we are of two minds. It presumes a clash of views between persons or within one person, and seeks to convince others, or ourselves, of the reasonableness of the principles upon which our claims and judgments are founded…justification proceeds from what all parties to the discussion hold in common…thus, mere proof is not justification…proofs become justification once the starting points are mutually recognized, or the conclusions so comprehensive and compelling as to persuade us of the soundness of the conception expressed by their premises…[C]onsensus…is the nature of justification (Rawls 1999, 508-509).

Only in the context of proffering, refusing, and accepting reasons does justification have meaning. However, justification is not merely the activity of proffering and rejecting reasons but the activity of doing so with others. Argument, says Rawls, presupposes some basic foundation of agreement – at the very least – to the activity of engaging in argument, to the activity of purveying and considering reasons in order to reach a justified endpoint.

It may be objected that this reading of Rawls places unwarranted emphasis on the social context of RE, rather than the process of articulating, weighing, balancing, and selecting information. Yet as Rawls makes clear, this is incorrect. Given his description of the ideal RE procedure, it is legitimate to interpret Rawls as holding that mere personal reflection is insufficient for ethical justification: justification requires the consideration of such a wide range

of information that it simply cannot be performed by singular reflecting individuals who do not interact in *dialogue* with others at least. The type of RE of concern to moral philosophy, says Rawls, occurs when "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." Yet he continues: "To be sure it is doubtful whether one can ever reach this state. For even if the idea of all possible descriptions and of all philosophically relevant arguments is well-defined (which is questionable), we cannot examine each of them. The most we can do is to study the conceptions of justice known to us through the tradition of moral philosophy and any further ones that occur to us and then to consider these (*ibid.*, 43). Thus Rawls' expresses skepticism regarding whether personal reflection alone justifies ethical theory. Whether reflective equilibrium is justificatory depends on the information the procedure begins with *and* incorporates over subsequent iterations. For Rawls, it appears to be an open question whether this information is of the sort that it is likely to be generated by a single individual, even one most familiar with moral and political philosophy. Perhaps it is. If so, then this would be a Herculean instance of personal reflection. Rawls suggests most mere mortals are incapable of adequate reflection upon such information. Thus it is appropriate to conclude that for him justification requires deliberation and dialogue to overcome natural limitations in human cognitive abilities.

5. Rawlsian Reflective Equilibrium

A close study suggests a distinct *Rawlsian reflective equilibrium* (RRE) that differs in important ways from other varieties. RRE takes moral principles, reflections on particular cases, and other information as inputs, and gives as output a justification for a specific higher-order concept. Moreover, on RRE these inputs are distributed across many persons, rather than located in a particular cognitive agent.

In addition to indicating that justification is a social endeavor, Rawls' account of reflective equilibrium also makes clear that it is a method for metaethics, rather than normative

ethics.[3] Thus, a successful iteration of the RE procedure would result in the definition of moral principles such as *justice*, as in "justice as fairness." By beginning with sharing information and ending with agreement about principles, justice, in this sense, is not justified for the purpose of guiding the subsequent actions of *individuals* who accept Rawls' analysis. Rather, accepting his analysis entails that one agrees Rawls' *definition* of justice is justified. The target of RRE is not how justice, so understood, ought to be applied in specific circumstances; its target is the justification of the principle.

6. The "Royal We"

An objection to this analysis could proceed merely by recognizing the *royal we*, the common practice of using "we" in scholarly writing in the place of "I." That is, one might charge that when using "we" Rawls' means "one," as in one person or individual. Consequently, Rawls' account of RE is just as logicistic and egocentric as Daniels' – RRE and RE are no different, and RE is a procedure an individual may use to bring an array of information into coherence, justifying resulting principles at semi-stable equilibrium.

This objection is not as important as it might first appear. Rawls' occasionally describes RE as a method carried out by an individual person, but there are also moments where Rawls' is ambiguous concerning whether he is referring to a single person or a group of deliberative agents. Yet at his clearest, Rawls expresses an interest in the collective will of a deliberating public and a commitment to capturing its justificatory force for moral theory. Consider his closing passage to the revised version of *Theory*. There he appeals to the vision he has in mind, of seeing ones place in society *"sub specie aeternitatis*," or "to regard the human situation not only from all social but also from all temporal points of view" (*ibid*., 514). Here, as elsewhere, Rawls returns to moral justification and consensus. What justifies is consideration of a

[3] This distinction may seem unnecessary but it is not. Ambiguity persists in the literature over the aim of RE. For instance, Daniels says RE aims at "theory acceptance or justification in ethics" (1979, 256), but he fails to distinguish between theories for guiding or evaluating actions and theories describing, *e.g.*, the meaning of moral discourse. Others more clearly believe RE aims at the former type of theory, such as Brandt, for whom RE justifies "practical principles for guidance of interpersonal relations and for evaluating plans of action" (1990, 26; cf. van Thiel and van Delden 2010).

sufficiently diverse range of considered judgments and moral principles. While it is possible that a single reflective individual may adequately perform the procedure, and while Rawls evidently recognizes that individual thinking is a part of the justificatory process, it is nevertheless clear that singular reflection will be the exception when justifying the meaning of principles, rather than the rule. What justifies is the shared agreement of many stakeholders, which will be best accomplished by many interacting individuals. There may be a "royal we" but there is no "royal road" to regarding the human situation from all social and all temporal points of view.

7. Three Virtues of RRE

There are three virtues to the interpretation of Rawls proposed here. First, it promises to provide an alternative framework for responding to objections against RE. Some argue that RE is either problematically circular or impracticable. In response, we may note that RRE requires the consideration of many individuals' judgments and principles in the process of justifying the meaning of moral concepts. Thus it does not rely on a single cognitive agent for instantiating the procedure; consequently, it does not place too heavy a computational burden on any given individual. Moreover, it is not problematically circular because it requires the consideration of many perspectives. As a social method of justification in metaethics, RRE calls for collaboration and critical evaluation among individuals to isolate disagreements and build consensus subsequent to an extremely broad process of consideration. On this view, to forge agreement on principles, in light of shared assumptions and shared descriptions of the state of affairs, is to define moral principles and hence to agree on the meaning of important moral terms.

This view of RRE recasts Rawls' position in contemporary metaethics, showing him to be sympathetic with those who see moral discourse as requiring recognition of, and engagement with, other individuals in order to justify fundamental principles (*e.g.*, Lovibond 1983; and to a lesser extent, Darwall 2006). As such, we should view moral theory "as the

attempt to describe our moral capacity" (Rawls 1999, 41), which requires deliberating with others about our firmest convictions as well as those that are least shared.

Perhaps, in this way, the charge of circularity could be further responded to by supplementing Rawls' metaethical commitments with a social moral psychology explicating the nature of shared judgments and their relationships to both individual moral cognition and consensus moral principles. Perhaps, that is, Rawls' account of a method of justification might supplement, or be supplemented by, the growing body of work on the shared nature of reasons, intentionality, and hence, justified group morality and action (*e.g.*, Laden 2012, Bratman 2009, Tuomela 1995). This, then, is the second virtue of this analysis: it promises to align Rawls' method with growing literature on the social foundations of rationality and cognition. Although this literature has received only modest attention in moral philosophy and ethics, it seems likely that as scholars pay more attention to the distributed nature of mind and reasoning they will also seek to explain the moral life by inquiring into its social manifestation.[4] By interpreting Rawls' account of reflective equilibrium as outlined here, his approach is shown to have far more affinity with this growing literature than could be appreciated on the received interpretation of his views.

Finally, at the very least, the interpretation of Rawls proposed here may serve as a corrective to the widespread tendency to interpret reflective equilibrium as a method of ethical justification carried out by a single individual, with particular emphasis on the logical consistency of sets of information, and directed at the justification of normative principles for evaluating action. Above, Rawls has been shown to emphasize information sharing and the bringing about of consensus in order to justify moral principles. Perhaps there are good reasons to doubt this interpretation, however, arguments need to be given for them. Thus, those interested in reflective equilibrium should see the value in clarifying Rawls' view, and thus, the value in the novel interpretation offered here.

[4] *E.g.*, Goldman (1999) on social epistemology, Hutchins (1995) and Clark (2008) on distributed cognition and extended mind, Bratman (2009) on shared agency and intentionality, and Laden (2012) on shared reasons.

Works Cited

Arras, John D. (2007). "The Way We Reason Now: Reflective Equilibrium in Bioethics." In B. Steinbock (ed.), *The Oxford Handbook of Bioethics*, pp. 46-71.

Beauchamp, Tom L. and James F. Childress (2009). *Principles of Biomedical Ethics, 6th Edition*. Oxford: Oxford University Press.

Brandt. R. B. (1990). "The Science of Man and Wide Reflective Equilibrium." *Ethics* **100:**259-278.

Bratman, Michael (2009). "Shared Agency." In C. Mantzavinos (ed.), *Philosophy of the Social Sciences*, pp. 41-59.

Clark, Andy (2008). *Supersizing the Mind: Embodiment, Action, and Cognitive Extension*. Oxford: Oxford University Press.

Daniels, Norman (1979). "Wide Reflective Equilibrium and Theory Acceptance in Ethics." *The Journal of Philosophy* **76:**256-282.

------- (1996). *Justice and Justification: Reflective Equilibrium in Theory and Practice.* Cambridge, UK: Cambridge University Press.

Darwall, Stephen (2006). *The Second-Person Standpoint*. Cambridge, Mass.: Harvard University Press, 2006.

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

Goldman, Alvin I. (1999). *Knowledge in a Social World*. Oxford: Clarendon Press.

Laden, Anthony S. (2012). *Reasoning: A Social Picture*. Oxford: Oxford University Press.

Lovibond, Sabina (1983). *Realism and Imagination in Ethics*. Minneapolis: University of Minnesota Press.

Haslett, D. W. (1987). "What is Wrong with Reflective Equilibrium?" *The Philosophical Quarterly* **37**:305-311**.**

Hutchins, Edwin (1995). *Cognition in the Wild*. Cambridge, MA.: MIT Press.

Rawls, John (1951). "Outline of a Decision Procedure for Ethics." *The Philosophical Review* **60:**177-197.

------- (1999). *A Theory of Justice*. Cambridge, Mass.: Harvard University Press.

------- (2005). *Political Liberalism: Expanded Edition*. New York: Columbia University Press.

Stitch, Stephen (1988). "Reflective Equilibrium, Analytic Epistemology, and the Problem of Cognitive Diversity." *Synthese* **74:**391-413.

Tuomela, Raimo (1995). *The Importance of Us: A Philosophical Study of Basic Social Notions*. Stanford: Stanford University Press.

van Thiel, Ghislaine J.M.W. and Johannes van Delden (2010). "Reflective Equilibrium as a Normative Empirical Model." *Ethical Perspectives* **17:**183-202.