Public Reason 1 (2): 12-30

Rawls: Construction and Justification

Stefan Bird-Pollan
Harvard University

Abstract. I examine Rawls' indebtedness to Kant in *A Theory of Justice*, Kantian Constructivism and in "Themes from Kant's Philosophy". I argue that the way Rawls develop the justification of *A Theory of Justice* relies heavily on Kant's claims that rationality requires reciprocity and that rationality is to be understood as moral rather than as instrumental. Rawls thus reveals something new in Kant's theory namely that for Kant the hypothetical imperative is actually subordinate to the categorical imperative. However, Rawls eschews Kant's attempt at proving that we are rational and thus committed to treating each other with respect, hence Rawls argument fails to show that we do, in fact, share the intuitions about justice as fairness that underlie Rawls' theory.

Key words: Rawls, Kant, morality, constructivism, justification.

The metaphysical problems that plagued Kant's deduction of morality in the *Groundwork* III have seemed, to many twentieth century philosophers who wanted to retain much of Kant's moral philosophy, so great that these contemporary thinkers have abandoned the attempt to ground pure practical reason altogether. The question I mean to pursue in this paper is whether a certain type of Kantian moral philosophy can get by without such a grounding. In Rawls one finds a writer who believes that much of Kant's ethical theory can be salvaged if one sidesteps the question of a metaphysical justification for morality and concentrates on the proceduralism necessary for justice.

The question, to put it another way, is whether the Kantian framework that Rawls adopts, lends itself to a non-metaphysical use. By this I mean that Kant's system may be metaphysical through and through and as such require the discharging of certain assumptions in its final form. This ultimate metaphysical assumption, I will argue, is that there is, in fact, not just a shared but a universal morality. The aim of this paper is thus to reconstruct the parallels between Rawls' argument and Kant's own, drawing out just how heavily Rawls leans on Kant to construct his theory. With this parallel in place, it will then be possible to determine whether, given the strong parallels I argue exist, Rawls' theory can still claim to be valid without working through the metaphysical assumptions Rawls explicitly rejects.[1]

Rawlsian constructivism is, as I hope to show, a worthy successor to Kant in the sense that it seeks to avoid the problems that have plagued generations of Kant interpreters – to find some way of making the categorical imperative 'work'. Rawls' strategy, by contrast, is to concentrate on the categorical imperative as a way of thinking about moral laws immanently, that is, as constantly articulated and enacted by the individual agent. For

1] It is interesting to note that several of Rawls' students have returned to the path of a metaphysics of sorts in order to ground the universality of morality. See, for instance, O'Neill 1996, 194, Herman 1993, 198 and in particular Korsgaard, 1996, 15; 2009, 189.

Rawls, the categorical imperative is just the mental process we engage in when we think about how to be just to other human beings. Rawls thus emphasizes respect for persons over moral psychology. Respect for persons entails that we treat others just as we want to be treated by others and this simply means, not seeking special treatment for oneself. Respect, Rawls argues, should (and generally does) enters into every thought about others. This type of thinking is modeled in both of Rawls' justifications for the liberal political society: the original position and the reflective equilibrium. The categorical imperative is a way of thinking which enables such respect for others.

I will argue, however, that, compelling though Rawls' interpretation of Kant's ethical theory is, its aim of presenting a non-metaphysical interpretation is only partially successful. Rawls is successful in giving a non-metaphysical account of reflection through the reflective equilibrium – a process in which each agent reflects on her considered beliefs and also takes into account the beliefs of others. Absent a universal (and therefore 'metaphysical') notion of practical reason which underlies such reflection, however, there is no way of showing that the conclusions of individual reflection cohere in any socially meaningful way. Indeed, this absence of cohesion is the result of Rawls' failure to take concrete suffering into account. By building his theory on the possibility of coherence between individuals, Rawls has, I will argue, sidestepped the problem of the perspective of justice altogether.

A further way of framing the issue is to see Rawls' and Kant's theories as objections to the egoist who believes that all she is committed to is taking the means to her ends, but not to anything further. In believing this, the egoist essentially resists the idea that the hypothetical imperative is framed by the categorical imperative or that the rational is framed by the reasonable. To refute the egoist one must, however, make precisely this move. And this move relies on the metaphysical assumption of our membership in a community which shares the same fundamental commitment to universal justice.

In reconstructing Rawls' thought, I will present the argument regressively, starting from Rawls' conception of autonomy and working backwards, always asking for a justification for the previous level of argument, until at last we arrive at the reflective equilibrium which is supposed to underwrite the whole conception of justice. The regressive reconstruction follows the argument Rawls gives in "Kantian Constructivism in Moral Theory", if not in *A Theory of Justice*, and underlines the acknowledged debt Rawls owes to Kant. The regressive argument also affirms that, after all, Rawls wishes to give a Kantian style grounding to his project since the regressive argument is itself a device used by Kant in order to arrive at a transcendental argument, and argument, I will argue, Rawls fails to deliver.

I. THE RATIONAL AND THE CATEGORICAL IMPERATIVE

In the interest of space, I will not spend much time on Rawls' twin concepts, the original position and the veil of ignorance. They both model what Rawls will call the 'ra-

tional' in "Kantian Constructivism". Suffice it to say that for Rawls, the original position is a regulative principle and thus a way of adjudicating between conflicting desires and inclinations.[2] The agent in the original position must be both autonomous and motivated by her reflection. She takes the means to her ends. That is to say, the original position must yield universally acceptable principles (as in the hypothetical imperative which, for Kant, is analytic) and it must ensure that these principles are acceptable to all.[3] The former condition is modeled in the original position by bargaining and the latter is modeled by the veil of ignorance.

Let us look at autonomy first. Rawls introduces the veil of ignorance to hide the parties' particular social and natural circumstances. The parties are asked to design a society without the knowledge about where they will be placed in the society, or which beliefs, moral, political or religious they will have.[4] All participants understand the basics of political affairs and economics and possess general knowledge. Thus they choose principles under which they are prepared to live, wherever they end up in society. The general social structure is just but blind to the particular inclinations of the agents. Under the veil of ignorance, just as in Kantian autonomy, we have no personal or particular sense of the good. We seek only justice, the ability to enjoy our particular notion of the good once we determine what that is.

Rawls also argues that there is a parallel between rational choice theory and the categorical imperative. Rawls says that the original position is in the tradition of social contract theory. Like the categorical imperative, it provides a way of responding to a practical problem: what ought I do? Rawls' two principles of justice are simply the moral law under the conditions of a modern liberal society, yielding more specific versions of the universal prescriptive of respect as stated in the categorical imperative.

We should note two points before we go on. In the model of the original position, Rawls has moved moral reflection from the first person perspective to public deliberation; from the 'I' to the 'we'. At least prima facie, the original position is not supposed to be all in the mind of one individual. The second point follows from the first. By changing the perspective of reflection from the first person to the third person, Rawls has also changed the moral psychology involved in accepting the outcome of deliberation. It is not clear that accepting the outcome of public deliberation has the same normative force as accepting the outcome of my own deliberation on the authority of the moral law.[5]

2] Rawls himself does not believe that Kant's categorical imperative actually provides a particularly good way of determining a content of the moral law. This is what his own theory of justice is supposed to provide (2007, 31).

3] Many have argued that a hypothetical agreement does not constitute a justification for the two principles chosen in the original position. See Nagel 1975, 114.

4] There has been considerable objection to the supposed neutrality made possible through the veil of ignorance. Onora O'Neill, for instance, notes that Rawls does not assume disinterest at all times during the original position process, but permits it with reference to the fate of future generations (1998, 121).

5] In a way, this is the problem Rawls will have to address in *Political Liberalism* where he will have to

Indeed, the central argument for the universality of morality hangs on this move from subjectively accepted norms to universally accepted norms. How is it, one might ask, that norms I develop for myself in my interactions with the world should be acceptable to all others? To put it another way, what is Rawls' argument against the egoist who believes that reasons are essentially private. Rawls seeks to address this issue in what follows.

II. THE REASONABLE AND THE RATIONAL

Rawls agues that underlying the original position and the application of the categorical imperative there is a conception of the moral character of the actors who reflect and thus abide by the moral law. In "Themes in Kant's Moral Philosophy" Rawls interprets these agents as both 'reasonable and rational'. Rawls uses these terms as a translation for Kant's *vernünftig*, which includes both senses. The two terms mark the distinction Kant makes between the two types of practical reason, pure and empirical practical reason. The former is found in the categorical imperative while the latter is exemplified by the hypothetical imperative. Rawls notes that Kant's conception of a person also marks the fact that, for him, the hypothetical imperative (empirical practical reason) is absolutely subjugated by the categorical imperative (pure practical reason) (1999a, 112). This is to say that the person who engages in moral reflection subjugates his rationally conceived maxims to the moral law.

Rawls characterizes his project in "Kantian Constructivism" as the attempt to: "establish a suitable connection between a particular conception of the person and the first principle of justice, by means of the procedure of construction" (37). This means that Rawls attempts to construct a philosophically coherent story about how the idealized conception of the person as reasonable and rational, can lead to a set of public institutions of justice we all can endorse. Before we examine what Rawls means by constructivism, we must understand what he means more exactly by the reasonable and the rational.

In political terms this means:

> [W]henever a sufficient basis for agreement among citizens is not presently known, or recognized, the task to justify a conception of justice becomes: how can people settle on a conception of justice, to serve this social role [of admissible social institutions], that is (most) reasonable for them in virtue of how they conceive of their persons and construe the general features of social cooperation among persons so regarded? (1999b, 305)

To put the issue slightly differently than Rawls does, we could say that the hypothetical imperatives each person at the bargaining table wishes to realize are limited by the recognition that each of the bargainers is equal and that it is thus unreasonable for one member to insist that the group agree to make an exception for that member. Thus the reasonable which models the demands of universality in the categorical imperative

---

show that we accept the results of the original position for reasons that are in a sense pure or moral rather than prudential. For a classic formulation of the objection to this move see Williams 1985, 205, ch. 4.

frames the debate about which particular hypothetical imperatives can be realized. The notion of universality, which Rawls interprets as equality, frames and restricts the particular rational plan of any actor. This turns classical liberal 'negative' freedom into a more communal 'positive' freedom. Thus when Rawls says that the original position is morally neutral, he means that there is no conception of the good involved in decision making itself. Morality, however, *is* in play in the sense that freedom and equality have a particular moral perspective, which is that the reasonable frames the rational.

But in order for the rationality of the original position to yield more than prudential agreement or a modus vivendi, Rawls must show that having a thin theory of the good allows each agent to move to a thick theory of the good. This is the point of introducing the distinction between the rational and the reasonable.[6] Rawls wants to show that instrumental reason as employed in the original position can be seen as an ethical capacity from a different perspective. This leads to a reinterpretation of the original position in "Kantian Constructivism", which relies more heavily on the notion of equality than its predecessor in *A Theory of Justice* did.

The movement from third person perspective to first person perspective occurs in three stages. It starts from rational autonomy (bargaining proper), moving to full autonomy (bargaining with reasonable or moral constraints) and finally ending up with the readers of Rawls' theory themselves (which finds its justification in the reflective equilibrium). What Rawls calls the rational or rational autonomy is modeled in pure procedural justice.[7] At the second stage, of full autonomy, Rawls adds to the conception of the person as free and equal two moral powers and two higher-order interests. The first power is that of having an effective sense of justice, the second is the power to form and revise and rationally purse a conception of the good. Corresponding to these are the higher-order interests of realizing and exercising these powers (1999b, 312).

The move to full autonomy and the reasonable, Rawls writes, is "expressed by the framework of constraints within which the deliberations of the parties (as rationally autonomous agents of construction) takes place" (1999b, 317). This framework is the reasonable ideal of fair cooperation. The framework, by which Rawls means the addition of the two moral conceptions of the person, reciprocity and mutuality, ensure that the plan of the good each person articulates for him or herself also includes the good of others. This is the doctrine of respect for persons as it is expressed in Kant's second formulation of the categorical imperative, the formula of humanity.[8] Here people are conceived of as an ends in themselves. Thus the two moral powers overlay the process of rational deliberation, transforming the instrumental deliberative process in the original position into a process of mutual recognition and fair cooperation. Rawls elaborates: "In justice as fairness, the

6] See Rawls 1993a, 503-4, and 1999b, 316. Also Baynes 1992, 122.

7] This means that the outcome is justified by if the means of arriving at it were just.

8] "Act so that you use humanity in your own person or in the person of any other, always at the same time as an end, never merely as a means." (Kant 1996, 58, Ak 4:429)

Reasonable frames the Rational and is derived from a conception of moral persons as free and equal. Once this is understood, the constraints of the original position are no longer external." (1999b, 319). I take this to mean that only the device of the original position (which models instrumental reason) imposes the constraint of fair cooperation on the people. For the people in the original position, social cooperation is not intuitive. But it is so for fully autonomous people who live in the institutions which the two principles of justice have helped to create. For they see themselves as possessing the two moral powers and thus restrict their pursuit of the good in the name of something more than the maximization of their material gain.

The movement of the two stages so far trades on the distinction between different perspectives. If we move back a little, we might recall that the purpose of the original position is to develop principles of justice out of our presuppositions about moral character. That is, what kind of laws would free and equal people come up with if left to their own devices? What Rawls does is to draw out first what free or rational individuals would do and then to overlay this with what people who are both free and reasonable would do. Rational people seek to maximize their benefit while reasonable people seek to maximize their benefit with the concerns of others in mind. This parallels exactly the structure that Kant argues for as well: we are rational beings insofar as we try to realize our ends by adopting the means to do so, but we are moral insofar as we adopt only those ends which we can will others to adopt as well.

Thus Rawls can say: "The unity of practical reason is expressed by defining the Reasonable to frame the Rational and to subordinate it absolutely; that is, the principles of justice that are agreed to are lexically prior to their application in a well-ordered society to claims of the good." (1999b, 319).

The lexical ordering of the reasonable over the rational also parallels Kant's division of practical reason into empirical practical reason and pure practical reason. While empirical practical reason– the hypothetical imperative– means acting according to any practical principle, pure practical reason– the categorical imperative– means acting according to the principle of the moral law.

However, there are still two elements missing from this argument. The first, to which we will now turn, is the question of how we get from the presupposed character of the agent as reasonable and rational to the content of the principle of justice, which so far has been described only formally. The second question, which we will come to after that, is what justifies the assumption of people as 'reasonable and rational' in the sense of being free to set their own goals. The second question comes down to what grounds Rawls assumption that we are, in fact, reasonable (or moral) and hence that I set *my* goals with other people's goals in mind.

III. CONSTRUCTIVISM

Constructivism is meant to be the way to get from a certain conception of the person (here, free and equal) to the appropriate principles of action for such a person. This means that constructivism seeks to draw out the content of the conception of the agent and to formalize it. That is, if the CI-procedure is the appropriate form of a rational principle, what is the appropriate material? The answer is the free and equal agent.[9] It is the answer to the question: what should we do when we act under the moral law or use our pure practical reason (which amounts to the same thing)?

In Kantian terms this means that: "the totality of particular categorical imperatives (. . .) that pass the test of the CI-procedure are seen as constructed by a procedure of construction worked through by *rational* agents subject to various *reasonable* constraints." (Rawls 1999c, 513-14). Each time we reflect and determine a law for ourselves we construct an element in a universal set of rules which can then be abstracted and turned into a general duty. Rawls' two principles of justice are a version of what might be arrived at in such an abstraction. The point, though, is that the maxims of conduct permitted or enjoined by rational reflection are not theoretical speculations; they are responses to actual needs for clarification of the permissibility of intended action.[10]

We can thus say that constructivism is the idea that the content of our highest moral principles stems from the rational and reasonable reflection upon our concepts as free and equal agents. Constructivism models autonomy in the sense that it constitutes the moral law or principle of justice from within its own rational and reasonable reflection. Nothing can count as a law for me without my having determined it for myself. This strongly echoes Kant's claim that there is nothing good in itself except the good will.[11]

Now, as before, there is here an emphasis on the first person perspective. That is, constructivism is just the CI-procedure insofar as it pertains to determining the content of the moral law. The content of the moral law has the content it has because I have (rationally) reflected upon it and have determined that it has this content. We must, however, keep open the possibility that when this first person perspective is switched to a third person perspective, as it is in rational choice, we loose normativity altogether. We will return to this issue.

Construction thus has two elements. First, it is a process internal to the agent and as such it is from a first person perspective. No one can reflect for me. Second, it is practical. Since reflection on the permissibility of performing an action stems from an incentive for action, the result of my reflection can only ever be manifested in my action itself. The

9] For this way of putting the problem see Korsgaard 1996, 123.

10] O'Neill notes that the constructivist position is anti-realist because it denies that moral facts are discoverable in theoretical terms. Constructivists believe that ethical principles are constructed by human agents, that these principles are practical and that they are objective. See O'Neill 2003,. and also Korsgaard 1996, 124.

11] See *Groundwork*, Ak 4:393.

result of my reflection can only ever be what I actually do, that is, what motivates me. If I say I ought to give $100 to charity and do not, I have actually decided to keep the $100. A practical constructivism thus relies on a notion of pure practical reason, that is, the idea that we are capable of reflecting on our ends by the use of the moral law or the two principles of justice.

Let us take a step back again and see where the argument has gotten us so far. Constructivism was introduced to provide a link between the conception of persons as free and equal (or as reasonable and rational) and the content of the principles of justice. Rawls' contention was that through the process of construction, or through autonomous reflection, these ideal agents would determine a set of principles which are able to govern the agents who have developed them in fair cooperation. Construction was then the way to bring out the content of the basic idea of the reasonable and rational agent without introducing any alien conceptions of how the world is or ought to be. The only tool available to the reasonable and rational agent in determining what the principles of justice are is reason. We also found that this constructivism proceeds from a practical point of view, which cannot be justified in theoretical terms. Justice is immanently constituted as doing that to which all involved have agreed.

Thus three of the four elements of Rawls' argument are in place. The original position has been established as yielding a universal principle. The presuppositions about the moral character (freedom and equality, reasonable and rational) of the agents who participate in the original position have been examined. And lastly, constructivism has presented a way for us to move from these presuppositions of moral character to the actual content of the formal characteristics of the moral law: the principles of justice.

The only element that is still missing is the justification of why we should think that we are actually those people in the original position who frame the rational by the reasonable. That is, what makes me think that I can assume that other people share the concern I have them. Where, in other worlds, does the universality or the reciprocity of the reasonable and rational framework lie. This is, to be clear, the basic assumption about morality that Kant was unable to provide in the *Groundwork* and an assumption which Rawls must make good on if the original postion, the reasonable as framing the ration and constructivism are to make sense.

IV. EXCURSUS: RAWLS AND KANT, PARALLEL ARGUMENTS

I have argued that Rawls has roughly followed the structure of the *Groundwork*. He has developed the categorical imperative in terms of the universality of the two principles of justice, corresponding to the first formulation of the categorical imperative.[12] Then Rawls has switched perspectives and has argued that the presupposition for such a law

12] "I ought never to act in such a way that I could also will that my maxim should become a universal law." (*Groundwork*, Ak 4:402)

is that people respect their humanity and, finally, Rawls has contended that in order to act under the moral law, we must imagine ourselves as instantiating the two principles of justice. For Kant this is the kingdom of ends. Let us examine these parallels in a little more detail.

Rawls writes: "In the first formulation [of the categorical imperative], which is the strict method, we look at our maxim from our point of view. (. . .) We are to regard ourselves as subject to the moral law and we want to know what it requires of us." (1999c, 505). This, I want to argue, is similar to the original position in which we want to know the formal structure a principle of justice would have.In the second formulation, however, we are to consider our maxim from the point of view of our humanity as the fundamental element in our person demanding our respect, or from the point of view of other persons who will be affected by our actions. Humanity both in ourselves and in others is regarded as *passive*: as that which will be affected by what we do (1999c, 505).

As I have already indicated above, I take this to be the perspective of drawing out the presuppositions about agents in the original position. To frame the rational by the reasonable means to see ourselves as passive in the face of the hypothetical imperative and to try to avoid damage to our humanity by restricting its scope. Our humanity is the material for the application of the CI-procedure in the sense that this is the purpose for that procedure. Rawls adds: "The point is simply that all persons affected [by my will] must apply [the CI-procedure] in the same way both to accept and to reject the same maxims. This ensures a universal agreement which prepares the way for the third formulation." (1999c, 505)

"In [the third] formulation we come back again to the agent's point of view, but this time we no longer regard ourselves as someone who is subject to the moral law but as someone who makes the law. The CI-procedure is seen as the procedure adherence to which, with a full grasp of its meaning, enables us to regard ourselves as legislators– as those who make universal public law for a possible moral community." (1999c, 506) This last formulation is clearly analogous to constructivism in the sense that in constructivism we develop positive law out of our conception of ourselves as free and equal.

I provide this juxtaposition of the structure of Rawls' and Kant's arguments not only to support Rawls' claim that *A Theory of Justice* is largely Kantian in orientation but to show that *A Theory of Justice* brings out central features of constructivism which must be seen as not just incidental but substantive contributions to Kant scholarship (§40). I further wish to argue that by tying his theory to Kant so closely, Rawls' theory is subject to many of the same difficulties as Kant's work. These difficulties have mainly to do with the problem of justification. For instance, the failure of Kant's deduction of morality has left Kant without a footing from which to say that humans are indeed able to interact respectfully with one another. Because Rawls avoids this push toward immanence and stays at what might be called the 'common sense' level, he also lacks a philosophically rigorous conception of intersubjectivity. Rawls' rejection of metaphysics, as I have said before, leaves him without an answer to the question of how people can actually be relied upon to treat each other with respect.

V. JUSTIFICATION AND THE REFLECTIVE EQUILIBRIUM

If the theory of construction is the justification for the two principles of justice, then what justifies construction? Rawls' answer, like Kant's answer to the problem of why humans should consider themselves free, is quite simply that constructivism is not justified in a theoretical way, but is given its authentication through cohesion into the perspective of existing humans who find that they agree with it. Justification is given through action. This notion of coherence is the final step in the three part development of authentication presented in "Kantian Constructivism".

Finally, Rawls comes to consider the last perspective, "that of ourselves– you and me – who are examining justice as fairness as a basis for a conception of justice that may yield a suitable understanding of freedom and equality" (for our own practical use) (320-21). Rawls continues:

> Here [in the third perspective] the test is that of general and wide reflective equilibrium, that is, how well the view as a whole meshes with and articulates our more firm considered convictions. (. . .) A doctrine that meets this criterion is the doctrine that, so far as we can now ascertain, is the most reasonable for us. (1999b, 321)

At this third perspective then, we have arrived at the criterion for a final justification of Rawls' theory. The problem for Rawls, as for Kant, is that we cannot prove that people believe themselves to be those ideal agents. Rawls is quite convinced that the failure of Kant's deduction of the moral law is sufficient to show that such an idealized theory approach makes no sense. So, according to the third perspective, the justification or authentication comes down to what Rawls calls the reflective equilibrium.

Let us now examine what the reflective equilibrium is in more detail. As Kenneth Baynes puts it: "reflective equilibrium refers to a condition in which an individual's concrete moral judgments have been brought into harmony with her higher-order moral principles" (1992, 69). This harmonization occurs first through a narrow process of reflective equilibrium in which one moves back and forth between concrete judgments (in, say, the manner of the categorical imperative in which the subject decides on a maxim and, using the categorical imperative procedure, determines whether it can be acted upon – if not, a new maxim must be created and tested) and then through the wide process of reflective equilibrium in which one's own judgments are bought into harmony with general social norms, shared by most readers of *A Theory of Justice*.[13]

Let us now turn to Rawls' own characterization of the process before turning to criticisms and defenses of this method. Rawls holds that his theory of justice describes our own sense of justice (1999a, 35). The justifications of his theory of justice, modeled by the original position and background conditions, are all reflections of our own considered

13] Rawls already characterizes the discussion about Justice as Fairness as taking place within a bounded society, one that endorses liberal democracy. Readers coming from outside this realm, may not agree with him.

judgments. The need to write *A Theory of Justice* in the first place, however, must have been generated by the knowledge that, on the face of it, not everyone currently *does* in fact share Rawls' conception of justice. The task of justifying the theory of justice thus must occur though a process of fleshing out those beliefs we actually all hold.[14]

The process of achieving reflective equilibrium systematizes our beliefs.[15] What we do in the narrow reflective equilibrium is thus similar to what we do in the CI-procedure. We take a practical problem which, admittedly, is more abstract than our everyday practical concerns, and reflect on it. For what is systematizing but bringing disparate concepts under a general principle of practical reason.

There is thus a positive and a normative side to the reflective equilibrium, or, as Thomas Scanlon put it, a descriptive and a deliberative side. As a method of arriving at an accurate portrait of justice, we must dig within ourselves to find normative notions we endorse (2003, 113). Both sides seem to be included in the following statement by Rawls: "we do not understand our sense of justice until we know in some systematic way covering a wide range of cases what these principles are." (1999a, 41). Indeed, in this statement of the purpose of the reflective equilibrium it is not possible to separate the two senses. Since, however, the process of the reflective equilibrium is a theoretical undertaking to which we subject our considered judgments, it seems appropriate to call it a method of deliberation.

The method itself is not explained in great detail in *A Theory of Justice*.[16] I will thus cite only the two main passages from this work in which Rawls describes the reflective equilibrium process as it pertains to original position:

> By going back and forth, sometimes altering the condition of the contractual circumstances [in the original position], at others withdrawing our judgments and conforming them to principles, I assume that 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. (18)

> A conception of justice cannot be deduced from self-evident premises or conditions on principle; instead, its justification is a method of the mutual support of many considerations, of everything fitting together into one coherent view. (19)

The reflective equilibrium begins with our considered opinions which must be made "under conditions favorable to deliberation and judgment in general". (40) A cen-

14] Who exactly the 'we' is, has been the subject of much debate. See, for instance, Okrin 1994, 125

15] See Rawls' formulation about the original position: "The conditions embodied in the description of the original position are ones that we do in fact accept. Or, if we do not, then perhaps we can be persuaded to do so by philosophical reflection." (1999a, 19)

16] Rawls' idea of the reflective equilibrium has been taken up in the fields of moral philosophy and the philosophy of science. See, for instance, the more rigorous formulation James Blanchowicz gives (which is not based strictly on Rawls' account). Likening the reflective equilibrium process to repairing a ship at sea, Blanchowicz writes: "It is not just the fact that one is resting on a dry part of the ship in one's efforts to repair a leaking part and that one may later rest on the repaired (formerly leaking) part to repair a new leaking (formerly dry) part that establishes genuine reflective equilibrium, but rather the fact that the ***way*** in which one rests on these respective parts is different in each case (. . .)." (1997, 126)

tral feature of these conditions, and one which will be relevant to criticism of the reflective equilibrium discussed below, is that the process of the reflective equilibrium is always subject-dependent. That is, it is always my judgment that comes in for consideration. In this sense, judgments are judgments only when they come with my reasons for the judgments attached.[17] They are thus not comparable to observational data.[18]

As a coherentist strategy, convergence in reflective equilibrium is only evidence of how much agreement we already have. It is not normative, in the sense that it might convince one who does not hold what I hold to change his or her mind. It is purely introspective. This is because, as Norman Daniels argues, coherentism in the form of the reflective equilibrium remains agnostic about whether there is any truth which it might approximate. This agrees with the point about anti-realism raised earlier according to which constructivism develops all 'truths' through reflection itself.

VI. PROBLEMS WITH THE REFLECTIVE EQUILIBRIUM

As we saw above, the method of the reflective equilibrium is a way of becoming clear about one's own ethical convictions. We examine our thoughts and our principles are measured against what we might have read or discussed with others. The important point to keep in mind is that we are now ourselves, comprehensive subjects with commitments to notions of the good. This means that each of us reflects from a different perspective.[19]

At this point we must, however, make a distinction which I mentioned earlier, namely the distinction between the content of the theory and its very possibility. For there is an ambiguity in the charge that we reflect from different perspectives as real existing agents. The charge might mean that, since we are different, we are not sure whether we will come to the same conclusions as Rawls does. But it might also mean that we would have completely different conceptions of morality or that morality might be denied altogether. The former point is addressed by the bulk of Rawls' argument while the latter point refers to a problem Rawls does not have much to say about.[20]

17] This is perhaps the place to note that Rawls never develops an adequate justification of the reflective equilibrium from the first person perspective, and thus ultimately leaves himself open to criticism from Kantians and others who regard the subject as the primary unity of ethical coherence. See the discussion of Christine Korsgaard and Onora O'Neill below. Both seek to remedy this deficiency in Rawls' account through their respective theories of practical reason.

18] See Daniels 1979, 12: 167-72. See also Brandt 1990, 128

19] This is what Sandel has in mind when he argues that the Rawlsian deontological subject is incapable of normative commitments because he or she has been cleansed of all contingency which would necessitate normativity in the form of judgment. In order to make the Rawlsian subject capable of normativity, normativity must be introduced at a later stage but this is impossible given the thinness of the subject as it is conceived in the original position (Sandel 1982).

20] Rawls addresses this issue in *Political Liberalism* where he talks about commitment to the liberal state as opposed to the modus vivendi, a temporary commitment which, in certain extreme cases, might seeks to overthrow the whole system.

The problem of different starting points for reflection brings with it a host of problems for Rawls. For instance, there is no longer any compelling connection between the perspectives of the different subjects being asked to endorse Justice as Fairness. The point is put nicely by Baynes who argues that there seems to be no reason for me to accept the results of the reflective equilibrium unless I am the one who has undergone the process myself. This, presumably, is what Rawls means when he writes that "each person has in himself the whole form of a moral conception" (1999a, 44). Here, whole must mean complete for me and not, as in Kant, universal. Thus, there does not seem to be any reason why I should be swayed by a subjective process of reasoning not my own.[21] We are thus back at the question of who the 'we' who endorses the considered moral judgments is and whether there is any connection among the individuals which make up the 'we'. In *A Theory of Justice* and in "Kantian Constructivism", this 'we' seems not to have been theorized at all where possible justification is concerned. This leaves open the possibility of egoism and hence the possibility that we do not, in fact, deliberate together as Rawls believes we do.

Scanlon, similar to Baynes, argues that the reflective equilibrium process is normatively underdetermined. This charge states simply that no conclusive evidence for or against Rawls' theory can be gotten from a coherentist justification.[22] Since the reflective equilibrium does not offer a determinate process by which one might arrive at ethical conclusions, it is quite possible for two people to start from the same premises and, using the reflective equilibrium method, still arrive at different conclusions. Rawls acknowledges this point when he says that his theory of justice is just 'a' theory of justice (1999a, 43-44).[23] But as a theory of justice it must include the claim that something is normative for us even if we cannot agree entirely on what it is. The problem is thus that a coherentist theory which seeks its justification in the reflective equilibrium is too weak to bind people of differing perspectives together because it cannot on its own overcome the differences that people with previous normative commitments bring to bear on their reflections. Coherentism, in other words, seems not to be able to provide consensus where there is none to begin with.

There is another, deeper objection here, however. Scanlon has argued that someone's employment of the reflective equilibrium commits the evaluator of the argument who undertakes it to nothing at all.[24] This question delves deeper since it asks the more

21] See Baynes 1992, 74.

22] Brandt argues, for instance, that Rawls' argument comes to a conclusion no more forceful than that: "A coherent set of beliefs can be made more convincing than another set even if there is nothing which can confirm or refute it." (1990, 272-73)

23] Concerning the intersubjectivity of the reflective equilibrium process, Rawls writes that the question must remain open: "I shall not even ask whether the principles that characterize one person's considered judgments are the same as those that characterize another's. I shall take for granted that these principles are either approximately the same for persons whose judgments are in reflective equilibrium, or if not, that their judgments divide a few main lines represented by the family of traditional [moral] doctrines (…)." . Rawls adds, referring to himself, that: "if we can characterize one (educated) person's sense of justice, we might have a good beginning toward a theory of justice". (1999a, 44)

24] See Scanlon 2003, 152 and O'Neill 1998b, 206-7.

fundamental question of whether morality exists at all and thus lays bare the assumption Rawls has so far been making about the reflective equilibrium, namely that it *is* the pure employment of practical reason. If the reflective equilibrium is, in fact, the pure employment of practical reason, there will be no problem with coherence beyond the merely technical problem of the correct assessment of the facts. We thus need some further argument about why the reflective equilibrium is, in fact, the employment of pure practical reason and not some other principle. This goes to the more fundamental question of the possibility of morality and thus relates quite clearly to Kant's own failed attempt at proving intersubjectivity.

The problem I am here insisting on is that the answer to the problem of the justification of the two principles of justice in the reflective equilibrium cannot be gotten through an analysis of the coherence of the two principles of justice with our own perspective as readers of political theory through the reflective equilibrium. The deeper problem suggested here turns on the question about the possibility of morality in general, which cannot be answered by coherentism precisely because it is a question of first principles or metaphysics, if you will. Indeed, coherentism can only give an evaluation of the rightness or justice of two principles of justice if it is assumed that coherence is really an expression of morality or practical reason.

Before taking up this final issue, we must look a little more closely at what the role of pure practical reason is in Rawls' theory. And this crucially depends on the perspective employed in the philosophical reasoning of *A Theory of Justice*.

VII. PURE PRACTICAL REASON AND THE FIRST PERSON PERSPECTIVE

We have now seen all four elements of Rawls' theory so I now want to take stock of the argument as a whole and make good on the promises for elaboration I made during the reconstruction of the argument. I will thus discuss what I see as the real problem in Rawls' ultimate justification of his theory, by which I mean the position of pure practical reason. In both the original position and the reflective equilibrium Rawls presents us with a conception of normativity, through bargaining and the interpretation of social norms, which seems to want to sidestep the question of the need for a justification of his claim for our ability to employ pure practical reason. I will argue, however, that a notion of pure practical reason must underlie both conceptions. I will then return to the issue of whether pure practical reason receives a foundation in Rawls' work.

The first problem I mentioned was the problem of what I argued was the substitution of the original position for the categorical imperative in *A Theory of Justice*. I noted that in this move Rawls replaced a first person perspective with a third person perspective. He seemed to be arguing that the process of deliberation under the veil of ignorance was just as good at leading to the two principles of justice as solitary reflection. Indeed, the substitution rather suggests that Rawls thinks rational choice is a better model for ethical thought than solitary reflection.

From a Kantian perspective, however, this move seems highly suspect. For what gives rise to normativity in Kant is that I make the law for myself, that I am an autonomous actor. As is clear from the reading I gave above, Rawls also considers this to be the case with the agents in the original position behind the veil of ignorance. But rational reflection, as Kant sees it, operates only from the first person perspective. That is, something is normative for me because I choose to adopt it as a principle. No one else can make me adopt as my end something I do not freely choose as an end. You may force me to do it, but it will not be my end.

This is just the familiar point that practical reason cannot be given a theoretical explanation. No one can convince me by argument that I should adopt their reasons. I must convince myself. So, if deliberation in the original position is really a 'compromise' as Rawls states, then the agreement reached in it is not normative for anyone since it does not represent a principle anyone actually endorses. The principle that has arisen through the compromise might, of course, still be adopted, but Rawls has not given us any argument for why those in the original position should adopt the principles they have reached in negotiation (1999a, 104-5).

There is a way out of this argument, of course. It is essentially that the original position with its multiple parties is just a way of representing what goes on in rational reflection in the CI-procedure. The move from the CI-procedure to the original position is just heuristic.[25] That this is so becomes quite clear, I think, when one examines the notion of constructivism, which is meant to connect the two principles of justice to the original position. Constructivism seeks to draw out the consequences of our presuppositions about the agents negotiating in the original position. But in order for us to be able to draw out anything about them, we must assume that they have something in common, namely the concepts of freedom and equality. This is why Rawls refers to these agents as idealized. In order for the process of construction to yield anything at all, 'idealized' must mean that they are at least generally the same. If this is so, then the move from free and equal individuals through construction to the two principles of justice merely mirrors Kant's movement from the *vernünftig* individual though rational reflection to the moral law.

As such, it is no mystery that the agents in the original position can come to a 'compromise' which is normative for all. The compromise is no compromise, it is really the presupposition of the moral theory underlying the make up of the agents – justice as fairness. There has thus been no shift from the first person perspective which admits of the use of practical reason to the third person perspective. There is also thus no issue of convincing anyone of the rightness of the two principles of justice.

So, as I think I have shown, the problem of normativity of the two principles of justice does not arise at the level of the original position since, fundamentally, the original position models the use of pure practical reason by an autonomous self. This does not mean, however, that the problem of normativity has been laid to rest. The normativity of the two

25] See Dworkin 1975, 129

principles of justice is simply moved back to the reflective equilibrium and to the question of its acceptance by comprehensive subjects. It also does not mean that the problem of the first person to third person switch and the problems this entails has gone away.

According to the arguments I have just given, we must conclude that Rawls' attempt to build greater stability for his system, through both the notion of bargaining in the original position and through the idea of wide reflective equilibrium which ultimately rely on a notion of public reason, is really reducible to the employment of pure practical reason by each individual. And the existence of pure practical reason in finite human beings is, of course, what Kant was unable to show in the deduction in *Groundwork* III.

VIII. THE PROBLEM OF NORMATIVITY AND THE NECESSITY OF ITS JUSTIFICATION

The fact that I have agued that the social anchoring that Rawls wants to give his theory by embedding it in broad social views is inconsistent does not mean, however, that the theory must be rejected or even that its steps are incoherent. I have merely shown that Rawls actually sticks far closer to Kant's general argumentation than is usually supposed. Two theoretical moves have been rejected but as long as we interpret these moves as merely heuristic, the general theory remains intact. It is thus time to come to the question of the final justification of the reflective equilibrium, in other words, whether there is an account of pure practical reason in Rawls' theory.

And here we come to the central problem of Rawls' justification. Kant saw his theory as hinging on the proof or authentication of the necessity of freedom and morality both in the deduction and in the fact of reason doctrine. Rawls does not think his theory requires such a grounding.

This brings us again to the problem of the first person and third person perspective of practical reason. I argued first that the agents in the original position, as autonomous and idealized, must share the same conception of freedom and equality, and that this means that they are really not substantially distinct in a way that would necessitate a compromise in determining the two principles of justice. Then I argued that the reflective equilibrium process which we must all engage in, in order to determine whether we actually believe ourselves to be similar enough to the idealized agents in the original position to endorse the two principles of justice they determine, also had to stem from a first-person reflection. Thus the claim and its authentication both stem from a first-person perspective.

Without a proof for the necessary identity between the results of the original position and the results of our own reflection, the most that can be said of the two principles of justice is that they cohere. And this is all Rawls wants to say. Rawls refuses Kant's deduction of morality in favor of Kant's fact of reason. And Rawls interprets the fact of reason as a coherentist justification for the two principles of justice.

Thus Rawls writes:

> Pure practical reason is authenticated finally by assuming primacy over speculative reason and by cohering into, and what is more, by *completing* the construction of reason as one unified body of principles: this makes reason self-authenticating as a whole. (1999c, 523)

The idea here is that since there can be no theoretical proof of freedom and morality, the only justification for morality that can be given is that we recognize ourselves as moral beings, that is, we recognize ourselves as the agents who participate in the original position.[26] This means, as we saw, that the speculative part of his theory, the original position and the mutual regard of rational autonomy, cannot be justified except through empirical endorsement by fully autonomous actors, the people engaged in ethical reflection – you and me.

Thus, much rides on the notion of recognizing ourselves in the idealized agents of the original position. This recognition comes down to believing ourselves to be capable of employing pure practical reason. And this belief is what Rawls means by cohering into a much broader notion of reason. Rawls thus offers a minimalist authentication of the possibility of morality itself.

I mentioned earlier that Rawls thought that the categorical imperative offered only modest help in determining contentful principles of justice. We are now in a position to make better sense of this claim. Because of his coherentist justification of the two principles of justice, Rawls does not maintain that *his* two principles of justice are the only ones possible. That is, he does not maintain that he has determined the precise content of the laws our social organization should take. He has proposed 'a theory' of justice which is open to revision.

But this is not a significant departure from Kant, since Kant was not proposing a significant content to the moral law. He was only interested in showing that there is such a thing as the moral law. This, however, is a position in which Rawls must follow Kant, since in order for there to be any kind of theory of justice, the possibility of a theory of justice must be given. And this is what Kant's deduction and later his doctrine of the fact of reason seeks to show.

The second claim is deeper, for it contains a thesis about the ultimate justification of morality. Here Rawls just assumes that the principle of practical reason really exists. In this sense, the revisability of the two principles of justice which, as we saw, are supposed to be derivatives or incarnations of Kant's categorical imperative, depends on there being such a thing as the categorical imperative or freedom in the first place. By seeking to give a weaker interpretation of the categorical imperative in terms of the CI-procedure, Rawls has given up on Kant's claim that the weak autonomy thesis must be turned into a strong autonomy thesis. Rawls has, in other words, given up on the idea of showing that humans are rational beings and has just assumed that we are.

26] See also O'Neill 2003, 356-57.

But giving up on the strong autonomy thesis means that, as I have argued, there is no answer to the question raised by Kant, namely, why should we think that we appetitive humans are motivated by rational laws and hence, why should we think that what you think is 'rational' is not just a way of subjugating me. This is one possible objection the egoist might make against Rawls. The whole question of justice, in other words, rests on showing that we are all in possession of a common rationality which can help us to overcome our appetitive natures and adhere to derivatives of the categorical imperative as Rawls or anyone else proposes them. If the possibility of rational agency is not circumscribed by reasonable agency, then the egoist will not be refuted.

To put the question one last way in terms of Kant's analytic and synthetic distinction: it might be analytically true that humans would follow something like Rawls' two principle of justice if they were rational, but to show this goes beyond the scope of Rawls' book. I hope, however, to have raised the issue of the grounding of reason with sufficient urgency to show that metaphysical neutrality is not an option for a theory of ethics.

pollan@fas.harvard.edu

REFERENCES

Baynes, K. 1992. *The Normative Grounds of Social Criticism: Kant, Rawls, and Habermas.* Albany, State University of New York Press.

Blanchowicz, J. 1997. Reciprocal Justification in Science and Morality. *Synthesis 110*: 447-68.

Brandt, R. 1990. The Science of Man and Wide Reflective Equilibrium. *Ethics 100*: 71-90.

Daniels, N. 1979. Wide Reflective Equilibrium and Theory Acceptance in Ethics. *The Journal of Philosophy 76*: 256-82.

Dworkin, R. 1975. The Original Position. In *Reading Rawls. Critical Studies on Rawls' Theory of Justice,* ed. N. Daniels. Oxford: Blackwell.

Herman, B. 1993. *The Practice of Moral Judgment.* Cambridge, Massachusetts: Harvard University Press.

Kant, I. 1996. *Groundwork of The Metaphysics of Morals.* In *Practical philosophy.* Cambridge: Cambridge University Press

Korsgaard, C. M. 1996. Kant's Formula of Humanity. In *Creating the kingdom of ends.* Cambridge: Cambridge University Press.

———. 1996. Reasons We Can Share. *Creating the kingdom of ends.* Cambridge: Cambridge University Press.

Nagel, T. 1975. Rawls on Justice. In *Reading Rawls. Critical Studies on Rawls' Theory of Justice,* ed. N. Daniels. Oxford: Blackwell.

O'Neill, O. 1996. Towards Justice and Virtue. Cambridge University Press.

———. 1998a. The Method of A Theory of Justice. In *John Rawls. Eine Theorie der Gerechtigkeit*, ed. O. Höffe. Berlin: Akademie Verlag.

———. 1998b. Kantian Constructivism in Ethics. *Ethics 99* (4): 752–70.

———. 2003. Constructivism in Rawls and Kant. In *The Cambridge Companion to Rawls*, ed. S. Freeman. Cambridge: Cambridge University Press.

Okrin, S. M. 1994. Political Liberalism, Justice and Gender. *Ethics 105*: 23-43.

Rawls, J. 1993a. Themes in Kantian Moral Philosophy. In *Kant and Political Philosophy*, eds. Ronald Beiner and William James Booth. Yale University Press.
———. 1993b. *Political Liberalism*. New York: Columbia University Press.
———. 1999a. *A Theory of Justice*. Cambridge, Massachusetts: Belknap Press.
———. 1999b. Kantian Constructivism in Moral Theory. In *John Rawls: Collected Papers*, ed. S. Freeman. Cambridge, Massachusetts: Harvard University Press.
———. 1999c. Themes from Kant's Moral Philosophy. *John Rawls: Collected Papers*, ed. S. Freeman. Cambridge, Massachusetts: Harvard University Press.
———. 2007. *Lectures on the History of Political Philosophy*. Ed. Samuel Freeman. Cambridge, Massachusetts: Belknap Press.
Sandel, M. 1982. *Liberalism and the Limits of Justice*. Cambridge: Cambridge University Press.
Scanlon, T. 2003. Rawls on Justification. In *The Cambridge Companion to Rawls*, ed. S. Freeman. Cambridge: Cambridge University Press.