A Defense of the Kantian Interpretation

Stephen L. Darwall

Ethics, Vol. 86, No. 2 (Jan., 1976), 164-170.

Stable URL:
http://links.jstor.org/sici?sici=0014-1704%28197601%2986%3A2%3C164%3AADOTKI%3E2.0.CO%3B2-R

Ethics is currently published by The University of Chicago Press.

A DEFENSE OF THE KANTIAN INTERPRETATION*

Stephen L. Darwall
University of North Carolina

I

Oliver Johnson argues that Rawls's claim to a Kantian interpretation of his theory of justice is unsupportable.[1] Indeed, Johnson holds that the main lines of Rawls's theory are such that it is in profound conflict with Kant's at three basic points. He argues that the Rawlsian theory of justice has no place for, and conflicts with, the Kantian notions of autonomy, the categorical imperative, and pure practical reason.

I find the idea that there is a Kantian interpretation of Rawls's theory a very attractive one. And though I shall not give a sustained argument for this position, I will try to give a defense of this view against Johnson's criticisms of it.

II

Just what does Rawls mean by the thesis that his theory has a Kantian interpretation? I take the main force of this claim to be that "the original position may be viewed, then, as a procedural interpretation of Kant's conception of autonomy and the categorical imperative."[2]

The interest in this thesis is not merely to locate Rawls's views within the history of moral philosophy. If the thesis is correct, it provides the hope of a deeper justification for Rawls's principles of justice. Rawls's two main arguments for his theory are well known: first, that the principles best systematize the considered judgments about justice that one would be prepared to accept in "reflective equilibrium," and second, that the principles would be chosen from a position (the original position) which embodies constraints that would seem compelling from reflective equilibrium. Both of these "coherence" justifications for the principles of justice may seem ultimately unsatisfying for a number of reasons. To begin with, the method seems to presuppose a rather substantial agreement in our considered judgments about justice in reflective equilibrium. Second, this justifica-

*I am indebted to the editor of *Ethics* for helpful comments on an earlier version of this paper.

1. Oliver A. Johnson, "The Kantian Interpretation," *Ethics* 85 (1974): 58–66.

2. John Rawls, *A Theory of Justice* (Cambridge, Mass.: Harvard University Press, 1971), pp. 147–48.

tion of the principles leaves unanswered the deeper question of why one should be interested in justice, even if it is true that our considered judgments about it can be organized by the principles. That is, it does not imbed a theory of justice in a theory of practical reason.

The Kantian interpretation suggests that there may be a deeper justification for the principles—namely, that they would be chosen from a perspective which, since it is the "procedural interpretation of Kant's conception of autonomy and the categorical imperative," it is compellingly rational to adopt.[3] To be sure, the attractiveness of this line of argument would depend on being able ultimately to make out the Kantian connections between autonomy, the categorical imperative, and pure practical reason. But, at the very least, this claim suggests the direction that a deeper justification of his theory might take and is of interest for that reason.

Thus, to my mind, the substance of Rawls's invocation of Kant in support of his theory is that there is a Kantian justification for the constraints on choice of principles imposed in the original position. I will not, therefore, be concerned with other proposed dissimilarities between Rawls and Kant. For example, it may be, as Joe Hicks has argued, that the notion of a social contract plays a rather different role in Kant's theory of justice than in Rawls's theory.[4] Whether or not that is the case does not directly affect the question of whether there is a justification in terms of the Kantian concept of autonomy (and related notions) for adopting the perspective embodied in the original position and accepting principles which one would choose from it.

III

Johnson's criticisms proceed "under three closely related heads, corresponding to the central concepts that Rawls believes himself to share with Kant; namely, (1) autonomy, (2) the categorical imperative, and (3) rationality."[5] His strategy is to argue, in each instance, that Rawls's claim that the respective Kantian concepts can be used to interpret his theory of justice is a mistake. Since on Kant's view there are fundamental connections between the three concepts in question, and since, therefore, Johnson's arguments against Rawls under each of these heads are

3. Notice that it is the coherence method of argument to Rawls's principles which Joe Hicks centers on as distinctly unKantian in "Philosophers' Contracts and the Law," *Ethics* 85 (1974): 20–21: "Rawls's manner of explication appears more economic: a quasi-bargaining process which weighs the available resources, as to regulative principles, against the objectives, as to ordinary judgments to be explained, and seeks the maximizing balance of the least expensive principles to secure the greatest richness of judgments." I think that Hicks is right to contrast Kant and Rawls on this point. Still, this is not directly relevant to the question of whether or not the thesis of the Kantian interpretation is valid. For, as I understand that thesis, it provides a different argument for, or explication of, the principles—namely, that they would be chosen from a position which is the "procedural interpretation of Kant's conception of autonomy and the categorical imperative." Thus, the question of whether or not the thesis advanced as the Kantian interpretation is true is independent of the sort of difference between Kant and Rawls that Hicks points out.

4. Hicks, "Philosophers' Contracts and the Law."

5. Johnson, p. 60.

at root the same argument, I propose to treat in depth only his argument with respect to autonomy and then to indicate how my points would apply to the other two concepts.

IV

Johnson's argument with respect to the Kantian notion of autonomy is that it is a mistake to think of principles arrived at from the original position as autonomous rather than heteronomous principles. This is a mistake because the principles are chosen by the parties in the original position in order to promote the interest of each. As long as decisions in the original position are so motivated, such decisions are necessarily heteronomous, since they spring from interest rather than respect for the practical law as such.

It is easy to be misled here. One may think that in order for principles arrived at from the original position to be autonomous principles (or laws of freedom as Kant calls them), the choice from the original position must itself be an autonomous choice. And it seems that as Rawls conceives the choice in the original position it is not an autonomous choice. In particular, the parties are conceived as choosing principles on grounds of interest, given the constraints of the veil of ignorance. Thus one is led to Johnson's conclusion.

But to be so led one must accept the initial premise, and I wish to argue that there is no good reason for doing so. It may well be the case that the choice of principles in the original position is a heteronomous choice because it is an interested choice and still be true that the decision of actual rational beings, not in the original position, to act under such principles is an autonomous decision, and hence that action on such principles is autonomous. Even if it is true that if one were under the constraints of the original position (most importantly, the veil of ignorance) one would want a particular principle adopted in one's own interest, it by no means follows that all, or even any, rational beings as they are actually placed in the world would want that same principle adopted in their interest. Thus, if a rational being chooses to act on principles which would be acceptable to him if he were under the veil (on the grounds that they would be acceptable to him under the veil), such a choice is by no means a choice on the basis of his interests and thus is not, on those grounds, a heteronomous choice.

Interestingly enough, Rawls seems to have anticipated the confusion which underlies Johnson's argument, and he explicitly warns against it:

> Since the persons in the original position are assumed to take no interest in one another's interests, . . . it may be thought that justice as fairness is itself an egoistic theory. It is not, of course, one of the three forms of egoism mentioned earlier, but some may think, as Schopenhauer thought of Kant's doctrine, that it is egoistic nevertheless. Now this is a misconception. For the fact that in the original position the parties are characterized as not interested in one another's concerns does not entail that persons in ordinary life who hold the principles that would be agreed to are similarly disinterested in one another. Clearly the two principles of justice and the principles of obligation and natural duty require us to consider the rights and claims of others. And the sense of justice is a normally effective desire to comply with these restrictions. The motivation of the persons in the original position must not be confused with the motivation of persons in everyday life who accept the principles that would be chosen and who have the corresponding sense of justice. . . . [Such an

individual] voluntarily takes on the limitations expressed by this interpretation of the moral point of view.[6]

Rawls's claim is that if one is willing to act only on the basis of principles acceptable from the original position, then, and only then, is one acting autonomously. To secure this claim one would have to be able to make a connection between being willing to act only on principles which one would will qua rational being (that is, to act on practical laws—principles which would be followed if "reason had full power over the faculty of desire")[7] and principles which one would find acceptable from the original position. This is the crucial connection to be made, since Kant understands autonomy in terms of the capacity of the will (as pure practical reason) to be a law to itself.[8] It is this connection which Rawls expresses by saying that the original position may be seen as "a procedural interpretation of Kant's conception of autonomy . . ." and on which, therefore, a Kantian justification for adopting the constraints of the original position would depend.

v

That such a connection can be made is at least suggested by the following commentary by Kant on the so-called realm of ends formulation of the categorical imperative:

> The concept of each rational being as a being that must regard itself as giving universal law through all the maxims of its will, so that it may judge itself and its actions from this standpoint, leads to a very fruitful concept, namely, that of a *realm of ends.*
>
> By "realm" I understand the systematic union of different rational beings through common laws. Because laws determine ends with regard to their universal validity, *if we abstract from the personal difference of rational beings and thus from all content of their private ends,* we can think of a whole of all ends in systematic connection, a whole of rational beings as ends in themselves as well as the particular ends which each may set for himself.[9]

The point here is that to will something as a practical law is to will it as a principle governing the behavior of all rational beings and hence to will it as a common law for all rational beings. Thus, one constraint on what one can will as a practical law is that one be capable of regarding it as a principle which could be willed by all other rational beings also.

Kant suggests that we can arrive at such a conception of rational beings under common laws only if we abstract from their own private ends and focus on what all would will as rational beings. Rawls's point is that the device of the original position, utilizing the veil-of-ignorance constraint, provides a methodological tool for performing such an abstraction. It allows one to derive what rational beings would will as common universal principles by forcing them to abstract from idiosyncratic differences between them.

6. Rawls, pp. 147–48.

7. Immanuel Kant, *Foundations of the Metaphysics of Morals,* trans. Lewis W. Beck (Indianapolis: Bobbs-Merrill Co., 1959), p. 17n.

8. Ibid., p. 65.

9. Ibid., p. 51 (emphasis added, except for the phrase "realm of ends").

VI

Johnson's rejoinder at this point would be that even though the parties in the original position operate under constraints which force such an abstraction, still the choice in the original position is one which is motivated by interest, and thus egoistic. Two things can be said in reply to this.

First, as I have argued, even if the choice in the original position is egoistic, it by no means follows that the willingness of actual rational beings to act only on principles acceptable from the original position is in any sense egoistic and thus heteronomous. Second, it is arguable that what is in one's interest under the constraints of the original position (in particular, under the constraint of the veil of ignorance) is in one's interest as a rational agent and not merely in one's interest in virtue of some desire one happens to have ("as belonging to the world of sense under laws of nature").[10]

The root idea here is that rational agency itself (or, in any case, rational human agency) requires the having of certain goods. A modicum of health, education, liberties, and wealth are necessary for one to exercise agency at all. Indeed, it is arguable that Rawls's primary goods are best thought of as goods from the point of view of rational agency, goods vital to one's existence as a rational agent. So much is entailed by Rawls's claim that "These [primary goods] are things that it is rational to want whatever else one wants. Thus given human nature, wanting them is part of being rational."[11] These goods are goods not just from the point of view of this or that particular end but from the point of view of one's having any ends at all—that is, from the point of view of one's being a rational agent.[12]

If this is correct, then the charge that the choice in the original position is an interested choice loses much of its bite. For it can now be conceived of as a choice from the point of view of one's interests as a rational agent. And thus it is arguably connected to what one would will as a rational agent, abstracting from one's own idiosyncratic desires, conception of the good, social position, etc.

VII

A rather important caveat is in order here. Although the veil of ignorance forces an abstraction from specific information about oneself (including one's

10. Ibid., p. 71. That a decision is motivated by an interest does not entail that it is heteronomous for Kant. After all, Kant entitles this section "Of the Interest Attaching to the Ideas of Morality." A choice can be autonomous and still be motivated by an interest (for example by an interest in morality) though not by an interest in some "external condition."

11. Rawls, p. 253.

12. It is important to note that the idea of primary goods is implicit in some of Kant's remarks also. One instance occurs in his discussion of the third example following the initial formulation of the categorical imperative in the *Foundations of the Metaphysics of Morals*. There he offers the following as support for the claim that a person could not will that everyone (or even that he himself) not develop his talents: "For, as a rational being, he necessarily wills that all his faculties should be developed, inasmuch as they are given him *for all sorts of possible purposes*" (p. 41, emphasis added). John Rawls reminded me of this passage. I am also indebted to Arthur Kuflik and Allen Buchanan for discussions of the view that primary goods are good from the point of view of rational agency. The idea is worked out in greater detail in a dissertation by Buchanan, "Autonomy, Distribution and the State" (Ph.D. diss., University of North Carolina, 1975), pp. 45–94.

desires and interests), there is a great deal of more general information that one has about oneself over and above the fact that one is a rational agent. Thus, one will know that one is a human being in the circumstances of justice. Since one has general knowledge of a psychological and sociological sort available to one, one will know in more or less detail interests, desires, and needs that one will have as a human being in such circumstances. Such information would be relevant to a choice of principles in one's own interest in the original position. Furthermore, one is to decide on principles for the basic structure of a society where this presumably means a group of human beings occupying some particular geographical area (though one knows nothing about that area in particular). All of this significantly restricts the generality of the choice from the original position. For even though one is forced to abstract from desires, interests, and features which would be idiosyncratic within the class of human beings in the circumstances of justice, one will have, as information relevant to a choice of principles, a great deal of information which may be idiosyncratic to human beings (in the circumstances of justice) within the potentially larger class of all rational beings. How does this effect any line of argument to the principles of justice on Kantian grounds?

It may well be the case that the principles of justice (even if arguable to from the original position) are not practical laws in the sense of laws which are valid for all rational beings. Still, if we conceive of principles of justice as principles which the basic structure of a human society ought to realize in the circumstances of justice, then the Kantian argument may still go through. For clearly the general information about the circumstances of justice and about human beings will be directly relevant here. The principles of justice may not possess universal validity in the sense of being valid for all rational beings, even though it is the case that they are valid for rational beings who are human beings placed in the circumstances of justice. Thus, were it true that the original position forces an abstraction from everything but information about oneself as a rational human being in the circumstances of justice, then it would be arguable that if there were principles which would be chosen by anyone so situated (that is, any rational being under those conditions), then such principles would be practical laws for rational beings under those conditions. The Kantian argument to the principles of justice would then be that they would be willed by any rational human being in the circumstances of justice if he were to attend to only those general features which characterize him as such a being, and hence would be willed by all such beings as common principles. Since the validity of such principles arises out of features of oneself qua rational agent (subject to the constraints of being human and in the circumstances of justice), they could be characterized as autonomous principles in that sense.

VIII

To recapitulate the argument: (*a*) Even if the choice from the original position is a heteronomous choice, it by no means follows that the decision to act only on principles acceptable from the original position (and hence action on such principles) is not fully autonomous. (*b*) It is arguable that the methodological device of the original position gives an interpretation to the Kantian idea of willing something as a practical law, a common law in a realm of ends subject to

the constraints of being human and in the circumstances of justice. Indeed, so much is hinted at in parts of the Kantian text itself.

IX

Johnson's arguments against Rawls's use of the Kantian notions of the categorical imperative and pure practical reason are similar to his objection to the Rawlsian use of the notion of autonomy. In each case, Johnson makes the mistake of supposing that since something characterizes the parties' choice of principles within the original position (and hence their grounds for accepting the principles), it must therefore characterize the principles themselves or the grounds of any actual person (outside the original position) for holding the principles. In the first instance, Johnson argues that the choice of the principles in the original position is conditional on the desire of the parties for primary goods, and as such the principles are mere hypothetical imperatives, conditional on one's having such a desire. Two things can be said about this. First, if someone accepts the principles of justice on the ground that he would choose them if he were in the original position (behind the veil, etc.), he does not accept them on the ground that the basic social structure's satisfying the principles is most likely to provide him with the highest index of primary goods as he is actually placed outside the veil. Thus, the principles are not hypothetical in that sense. Furthermore, the desire for primary goods is not merely one desire among others. It is arguable that it is a desire which is preeminently rational for one to have, given that one is a rational human being in the circumstances of justice.

With respect to the Kantian notion of pure practical reason, Johnson argues that since the parties within the original position are characterized as (*a*) mutually disinterested and (*b*) rational in the economic sense of choosing whatever will be most in their own interest, Rawls's notion of rationality is at odds with Kant's. Clearly Johnson has again misidentified the sense of 'rational' in which the parties within the original position are assumed to be rational with Rawls's notion of reason per se. Though Rawls is not terribly explicit about his conception of reason, in the final section of the book he alludes to a conception of rationality rather different than the narrow economic notion: "Within the framework of justice as fairness we can reformulate and establish Kantian themes by using a *suitably general conception of* rational choice."[13]

To be sure, one would want to emphasize the same caveat here as before. Principles which one would will from the original position are not categorical imperatives or practical laws in the sense that they are valid for the will of any rational being—though it is arguably the case that they are valid for any rational human being who happens to be in the circumstances of justice.

As I said at the outset, this paper is not intended to be a sustained argument for the Kantian interpretation. Much greater clarity is yet required about Rawls's views, Kant's views, and the connection between the two. Nevertheless, I hope that I have shown that such a program has not been rendered otiose by Johnson's criticisms.

13. Rawls, p. 584 (emphasis added).