Published in *Social Theory and Practice* 13, 1(1987) 49-76

Personal Continuity and Instrumental Rationality in Rawls' Theory of Justice

Adrian M.S. Piper

I want to examine the implications of a metaphysical thesis which is presupposed in various objections to Rawls' theory of justice.[1] Although their criticisms differ in many respects, they concur in employing what I shall refer to as the *continuity thesis*. This consists of the following claims conjointly:

(1) The parties in the original position (henceforth the OP) are, and know themselves to be, fully mature persons who will be among the members of the well-ordered society (henceforth the WOS) which is generated by their choice of principles of justice.

(2) The OP is a conscious event among others, integrated (compatibly with the constraints on knowledge and motivation imposed on the parties) into the regular continuity of experience that comprises each of their ongoing constitutes lives.

(3) The parties in the OP thus are, and regard themselves as, psychologically continuing persons, partially determined in personality and interests by prior experiences, capable of recollection and regret concerning the past, anticipation and apprehensiveness regarding the future, and so on.

Thus, for example, some early criticisms of Rawls' *Theory of Justice*[2] centered on what they took to be the individualistic assumptions embodied in the OP: Adina Schwartz argued that Rawls' assumption that the parties prefer a greater rather than a lesser amount of primary goods would contribute to a WOS based on a preference for more rather than less wealth, and that this condition would be unacceptable to one who discovered herself to be a socialist.[3] Similarly, Thomas Nagel[4] argued that the very concept of primary goods biases the choice of principles individualistically, against conceptions of the good that depend on the social interrelationships among individuals, and so may require the parties in the OP to commit themselves to a set of social arrangements that contravenes their deepest convictions once the veil of

---

[1] This discussion originated as a graduate seminar paper for Professor John Rawls in 1976, and I am grateful for his comments on it. I have also benefitted a great deal from criticisms of an earlier draft by Peter Dalton. It is excerpted from a longer manuscript in progress, *Rationality and the Structure of the Self*. Work on this manuscript was supported by a University of Michigan Rackham Faculty Fellowship and an Andrew Mellon Post-Doctoral Fellowship at Stanford University, 1982-84.

[2] (Cambridge, Mass.: Harvard University Press, 1971). All page references to this work will be in the text, preceded by *TJ*.

[3] Adina Schwartz, "Moral Neutrality and Primary Goods," *Ethics* 83 (1973): 294-307. See especially pp. 304-06.

[4] In "Rawls on Justice," *The Philosophical Review* 87 (1973): 220-34; reprinted in *Reading Rawls*, ed. Norman Daniels (New York: Basic Books, Inc., 1974).

ignorance is lifted. David Gauthier attacked Rawls' assumption of economic rationality, showing that parties guided by instrumental reasoning in the OP would choose, not principles to structure a society based on justice as fairness, but instead those that would structure a "private society" instrumental to the pursuit of their individual utility-maximization.[5] Finally, Richard Miller argued that an individual in the OP who turned out to have been a member of the ruling class with an acute need for wealth and power in the society preceding the OP would find her interests frustrated by the egalitarian requirements of the difference principle.[6] Each of these criticisms called attention to the possibility of a disparity between the interests or beliefs of the parties in the OP and the conditions they may confront in the WOS that is supposed to result from their choice. Hence each presupposes the continuity thesis.[7] More recent criticisms of Rawls' theory presuppose it as well.[8]

Although the continuity thesis as stated above is not at odds with any of the conditions that define the OP, its exegetical validity is a matter for discussion. I shall be concerned to argue that if it is indeed contained in or a consequence of Rawls' theory, then it casts into doubt the capacity of the OP to generate or justify any principles of justice at all. On the other hand, if the continuity thesis is viewed as dispensable and unnecessary to the Rawlsian enterprise, then Rawls is correct in maintaining the irrelevance of the question of personal identity to the construction of his moral theory.[9] In this case, the contract-theoretic, instrumentalist justification for the two principles of justice (henceforth the 2PJ) needs to be supplanted by a modified conception of wide reflective equilibrium. The considerations that form the bulk of this discussion then may be understood as providing a rationale for Rawls' recent revisions in the model of justification on which his theory of justice rests, and for his

---

[5] David Gauthier, "Justice and Natural Endowment: Toward a Critique of Rawls' Ideological Framework," *Social Theory and Practice* 3 (1975): 3-26.

[6] Richard Miller, "Rawls and Marxism," in Daniels, *Reading Rawls*.

[7] Considerations of space require that detailed textual arguments to support this claim be deferred to another occasion. I hope the intuitive point is obvious.

[8] See, for example, Anthony Kronman's and Samuel Scheffler's comments on Rawls' Tanner Lecture, "The Basic Liberties and Their Priority," *The Tanner Lectures on Human Values, Vol. III* (Salt Lake City: The University of Utah Press, 1982).

[9] John Rawls, "The Independence of Moral Theory," *Proceedings of the American Philosophical Association* 48 (1975): 5-22.

increasing emphasis on us as moral mediators between the OP and the WOS.[10]

Now I want to consider the question of whether or not, given the textual evidence, anything like the continuity thesis is stated or implied by Rawls, and what problems for his theory, if any, turn on a positive or negative answer to this question.

I

To begin with, there is much in *A Theory of Justice* to lend support to the continuity thesis. Certain passages on what Rawls calls the strains of commitment suggest that the parties in the OP are psychologically continuous with identifiable members of the WOS((1) and (2) of the continuity thesis). For example, when Rawls stipulates that the parties have a sense of justice in that they "can rely on each other to understand and act in accordance with whatever principles are finally agreed to. Once principles are acknowledged the parties can depend on one another to conform to them" (*TJ* 145), the importance of insuring that these individuals are, in the WOS, capable of adhering to the commitment they made in the OP is evident. This is re-emphasized later when Rawls asserts that:

> In view of the serious nature of the possible consequences of the original agreement, the question of the burden of commitment is especially acute. A person is choosing once and for all the standards which are to govern his life prospects… the parties must weigh with care whether they will be able to stick by their commitment in all circumstances (*TJ* 176).

These claims clearly presuppose that the parties in the OP are, and know themselves to be, psychologically continuous with particular members of the society the basic structure upon which they now decide. Also, in discussing sound procedures of moral education in the WOS, Rawls proposes that "in agreeing to principles of right the parties in the original position at the same time consent to the arrangements necessary to make these principles effective in their conduct." (*TJ* 515) This condition is clearly meant to insure the conformity to principle in the WOS of the parties in the OP. Finally, Rawls

[10] See in particular Lectures I and III of his Dewey Lectures, "Kantian Constructivism in Moral Theory: The Dewey Lectures 1980," *The Journal of Philosophy* 77 (1980): 515-72 and "Justice as Fairness: Political not Metaphysical," *Philosophy and Public Affairs* 14 (1985): 223-51.

makes clear in subsequent discussion[11] that the parties in the OP are to be conceived as members of a WOS.

Also relevant to the continuity thesis are those passages in *A Theory of Justice* which suggest that the parties in the OP are continuing persons in that they are partially determined in their tastes and values by events prior to the OP ((2) and (3) of the continuity thesis), and that this must be considered in the subsequent WOS. Rawls claims, for example, that the parties in the OP are to decide in advance the principles which are to regulate their interaction (*TJ* 11, 31), and further that they do this without a knowledge of their more particular ends:

> They implicitly agree, therefore, to conform their conception of their good to what the principles of justice, require, or at least not to press claims which directly violate them. An individual who finds that he enjoys seeing others in positions of lesser liberty understands that he has no claim whatever to this enjoyment (*TJ* 31).

The evidence here is strong that Rawls is canvassing the possibility that one such individual might discover, when the veil of ignorance is lifted, that her prior personal interests conflict with the principle she has chosen in the OP. Rawls' response is that such interests are simply to be disregarded. Further, his controversial claim that "it may turn out, once the veil of ignorance is removed, that some of [the parties in the OP] for religious or other reasons may not, in fact, want more of these [primary] goods" (*TJ* 142) provides additional support for the thesis that the parties in the OP are continuing persons, partially determined by their psychological histories, for whom the OP is an event among others in their conscious lives. This is because the implication here, as in the passage quoted above, is that the parties in the OP might subsequently discover in themselves psychological tendencies or desires that are in no sense determined by the decision made in the OP, hence must be determined by forces prior to that event. Nevertheless, those forces must continue to operate after it in order for the requisite discovery to be made. Again, the same point is made even more strongly later:

> How can the parties possibly know, or be sufficiently sure, that they can keep such an agreement?... any principle chosen in the original position may require a large sacrifice for some. The beneficiaries of clearly unjust institutions (those founded on principles which have no claim to acceptance) may find it hard to reconcile themselves to the changes that will have to be made. But in this case they will know that they could not have maintained their position anyway (*TJ* 176).

---

[11] "Reply to Alexander and Musgrave," *Quarterly Journal of Economics* 88 (1974); 633-39.

Finally, there are auxiliary passages which, when taken together, clearly buttress the conception of the continuity of the parties as identifiable both prior and subsequent to their participation in the OP. On page 166 of *A Theory of Justice*, Rawls observes that "The persons in the original position know that they already hold a place in some particular society;" and later, in discussing the strategic advantages of stipulating a four-stage sequence during which progressively more information is made accessible to the parties in order that they may make increasingly specific choices about constitutional and legislative matters, he says, "So far I have supposed that once the principles of justice are chosen the parties return to their place in society and henceforth judge their claims on the social system by these principles" (*TJ* 196). These two passages establish that the OP as an event is integrated into the continuing personal histories of the parties ((2) of the continuity thesis).

Further evidence of the continuity thesis is to be gleaned from the concluding paragraphs of the *Dewey Lectures*[12], where Rawls claims of the parties in the OP that:

> persons so conceived and moved by their highest-order interests are themselves, in their rationally autonomous deliberations, *the agents who select the principles that are to govern the basic structure of their social life* (*DL* 572; emphasis added).

Moreover, in this more recent discussion, Rawls frequently characterizes the parties in the OP in terms similar or identical to those that characterize the members of the WOS (see, for example, the description of each as self-originating sources of valid claims at *DL* 548, 564, and 543 respectively).

From all these claims jointly, we are quickly led to the conception of particular individuals, mature and partially formed by their own pasts and the previous conditions of their society, who voluntarily come together and, temporarily assuming the constraints and veil of ignorance of the OP, choose principles that are henceforth to govern their claims upon one another. The veil of ignorance is then lifted gradually, in accordance with the four-stage sequence (*TJ* Sec. 31). These individuals recover knowledge of themselves, their pasts, their habits, interests, and conceptions of the good, and immediately proceed to realize their chosen WOS in conformity with the 2PJ. This is the conception the continuity thesis expresses.

The truth of the continuity thesis gives credence to an implication common to each of the early criticisms of Rawls mentioned earlier. Schwartz and Nagel both claimed that someone of a strongly socialist or communitarian persuasion might be frustrated in her efforts to realize her conception of the good in Rawls' WOS. Gauthier argued that individuals with

[12] See Note 10. Page references will be in the text, preceded by *DL*.

the economic rationality Rawls ascribes to the parties in the OP would repudiate the 2PJ for others that better enabled them to pursue their individual interests. And similarly, Miller's criticism can be understood as suggesting that someone with a highly individualistic, ruling class-determined conception of the good might be frustrated in realizing it by choosing the difference principle. Now if the continuity thesis is true, the parties in the OP must at least consider the possibility that in fact they may be either of these types of people, in addition to numerous other possibilities (for example, that their conception of the good includes seeing other persons in positions of lesser liberty). They must consider the possibility that by choosing as they do -- however they choose -- they risk at the very least the extreme frustration of their deep-seated desires and conceptions of the good; or at most their gradual extinction as continuing personalities with identifiable tastes, interests, and values. Thus they must be prepared to give up all that is central to their prior sense of self for the sake of those principles which they agree are to regulate their behavior. But since they are also assumed by Rawls to be primarily concerned in the OP to advance their own interests and conceptions of the good in the subsequent society, it is difficult to see how both conditions can be satisfied; and how they can therefore choose principles of justice under the constraints of the OP at all.

II

But now suppose this dilemma is solved, as do Rawls' critics. Suppose, that is, that the truth of the continuity thesis is consistent with the parties' choice of some principles of justice that govern their society once the veil of ignorance is lifted. In this case, there are troubling implications both for Rawls' essentially instrumentalist justification of the 2PJ, and for the method of wide reflective equilibrium in which they are embedded. In *A Theory of Justice*, Rawls seems clearly committed to justifying these principles as necessary, instrumentally rational means for furthering the interests of free and instrumentally rational persons (*TJ* 129; also 432). In determining the basic structure of the WOS, these principles indirectly constrain the interests, aspirations and conceptions of the good of its citizens, and so the actions they take to realize them:

> In justice as fairness, persons *accept in advance* a principle of equal liberty… They implicitly agree, therefore, to conform their conceptions of their good to what the principles of justice require, or at least not to press claims which directly violate them (*TJ* 31; emphasis added).

Rawls' concept of the OP is thus designed to produce the choice of the 2PJ as an outcome of the parties' recognition in the OP that under conditions of

moderate scarcity of resources to which all have a *prima facie* equal claim, this is the most instrumentally rational way for each of them to secure their own interests in the resultant WOS (*TJ* 119; also see section 22). Now the continuity thesis also implies that from the point of view of any single individual in the OP, the 2PJ are necessary means to the realization of her system of ends under the circumstances of justice only if, when the veil of ignorance is lifted, and she discovers her own conception of the good, she has no cause to regret her choice of the two principles, no matter what that conception of the good turns out to be (*TJ* 421-22). But as we have already seen, Rawls himself acknowledges that this may not be the case (*TJ* 176). Moreover, the beneficiaries of unjust institutions also know that their inability to maintain their favored positions is a direct consequence of having chosen principles of justice which, when implemented, so alter the circumstances of their life that their original plans of life, that is, continuing to benefit from these institutions, become untenable. But we have also seen that if the parties know, in the OP, that implementation of the 2PJ may thus require them to sacrifice rather than advance their conceptions of the good in this way, they cannot fail to see the instrumental irrationality of choosing the 2PJ in the first place.

It may seem that the parties do not have reason to regret their choice if they recognize that the 2PJ were the best available alternative open to them.[13] And perhaps it is true that no alternative principles of justice would have the effect of securing their future happiness and security. But it was open to them not to choose principles of justice at all, in other words, to opt for some version of the "No Agreement Point" (*TJ* 147). Unless we assume that the parties were forced into the OP -- surely an unpalatable assumption in view of the parties' freedom and autonomy (*TJ* 11, 13), and Rawls' allegiance to traditional social contract theory, it is open to the parties to regret choosing to live by principles of justice in the first place, rather than to maintain the *status quo* in their previous society. Rawls does not consider the latter as a viable alternative for the parties in the OP, but there is no clear reason why he should not. For unlike traditional social contract theory, the parties do not, on the continuity thesis, enter into the OP from a state of nature mutually acknowledged as unacceptable. So it is consistent with the constraints on information expressed by the veil of ignorance (*TJ* 12, 136-37) -- that is, that the parties know nothing of the circumstances of their own society -- that the parties elect to take their chances in their society as it has been up to now, rather than risk having to abdicate everything that gives meaning and satisfaction to their lives, even if this requires the deliberate perpetuation of social injustice.

---

[13] I am grateful to John Rawls for discussion on this point.

For there is, in addition, nothing in the description of the parties' motivation, circumstances, or interests in *A Theory of Justice* that commits them to choosing just principles for society. By hypothesis, they are moved by the desire to further their conceptions of their own interests, or good, whatever this may turn out to be (*TJ* 129). And the fact that it is the circumstances of justice that move them to deliberate (*TJ* 128) does not imply that they must opt for just principles to adjudicate their claims. Rawls often talks as though the specification of the circumstances in which questions of justice arise naturally generates a motivation to seek a just resolution of the conflict of interests which characterizes the situation (*TJ* esp. 126-27). But there is no reason to assume this. An instrumentally rational individual whose interests are already immoral or unjust has no *prima facie* reason even to view those circumstances as circumstances of justice, for there may be more efficient ways of removing any obstacles to the achievement of her ends. In the face of such considerations, one's sense of justice might well remain dormant.

Thus for a continuing, instrumentally rational individual whose plans and projects already violate the moral constraints embodied in the 2PJ, these principles are not necessary means but rather obstacles to the achievement of these plans. Since all parties in the OP must canvass in advance the possibility of being precisely such individuals when the veil of ignorance is lifted, they must view these principles as expendable, just as are any other inefficient means to the instrumentally rational achievement of one's ends; and so reject them accordingly.

The general problem begins to emerge. In order to motivate the choice of principles of justice, Rawls must presuppose that the parties in the OP have not only a conception of the good they want to advance and a sense of justice, but in addition a motivationally effective desire for or interest in justice. Otherwise the 2PJ are not instrumentally rational for them to choose. Now in the *Dewey Lectures*, Rawls reformulates the parties' motivation in the OP in order to meet this requirement. There the parties are described as being moved by their highest-order interest of developing and exercising their sense of justice (*DL* 525-26). First, this means that whatever else they want, they know at least that they want to be moral persons. This reformulation thus purchases motivation to choose principles of justice at the cost of attenuating the opacity of the veil of ignorance. More seriously, it attenuates the ability of the OP to provide an independent, choice-theoretic justification[14] for the powerful moral conception the WOS expresses.

---

[14] Rawls states his commitment to this kind of justification at *TJ* 16, 47, 94, 119-21, 125, 172, 583. He retracts it at *DL* 572, and again more forcefully in "Justice as Fairness: Political not Metaphysical," p. 237, fn. 20.

To see this, consider the justificatory role of the circumstances of justice in determining the choice of the 2PJ in *A Theory of Justice*. Here the idea is that the OP represents the salient features of the situation in which equally positioned agents compete for scarce resources to further their unstated, mutually conflicting ends. These circumstances are compelling to us because we all understand what it means to have to compete for scarce resources in our daily lives. To show that these circumstances plus the other special conditions that define the OP generate certain principles of justice would provide a powerful incentive for us to accept them, for it would depict those principles as a rational outcome of conditions essentially reflected in our experience. But the reformulations in the *Dewey Lectures* require that each of the parties' conceptions of the good be constrained by their shared highest-order interests in developing and exercising their senses of justice. To stipulate at the outset that the parties are overridingly motivated to act on principles of justice is to ascribe to them a motivation that *a fortiori* overrides the motivational incentive of the circumstances of justice. In this case, all the parties must share a nonconflicting highest-order interest, namely their interest in developing and exercising their senses of justice. Moreover, this interest must be partially determinate, since they must therefore have some conception of what justice is, in order to be moved to develop and exercise their sense of it (to see this, try substituting "sight" for "justice"). But the relative determinateness of their conception of justice in turn undermines the sense in which they can be said to choose just principles through rational negotiation, consistently with the assumption of pure procedural justice (*TJ* 120, 136). It also undermines Rawls' claim, first made in *A Theory of Justice* (128) and developed more fully in the *Dewey Lectures* (557-60, 561, 564, 571), that the parties are not bound by antecedent moral ties. The more fully the sense of justice is conceived as a motivationally effective intentional object for the parties in the OP, the more motivationally otiose the subjective and objective circumstances of justice become.

Indeed, that the parties are overridingly motivated to realize and exercise the "capacity… to understand, to apply and to act from (and not merely in accordance with) the principles of justice" (*DL* 524) tautologically requires them to choose the 2PJ in order to realize this aim. Rawls has ensured that it is instrumentally rational for the parties in the OP to choose the 2PJ by stipulating in advance that that is what they are most highly motivated to choose, regardless of the further ends they serve.[15] That is, the 2PJ are, in the

[15] In fact it is less than completely clear whether "the principles of justice" in the passage just cited refer to the 2PJ, or to whatever principles the parties in the OP choose. The phrase occurs in the context of a description of the parties' moral powers

Dewey Lectures, no longer justified as instrumentally rational means to the promotion of a circumscribed range of conceptions of the good at all. They are stipulated to be final ends, part of each of the parties' conceptions of the good itself. This stipulation makes it much harder to conceive the OP itself as generating, rather than presupposing, the moral ties it was its original function to justify. Thus in order to ensure the instrumental rationality of the two principles of justice to the parties in the OP, Rawls has built the choice of the 2PJ into the OP in such a way as to nullify the issue of their instrumental rationality altogether. So either the parties must be viewed as motivationally uncommitted to making a specifically moral choice,[16] in which case a return to the *status quo* is a viable alternative; or else we must assume a priori that they are morally motivated to choose the 2PJ, in which case it remains an open question how those principles are to be independently justified. Elsewhere I try to show that this strategy is inevitable, if fatal, for the committed instrumentalist.[17]

---

and highest-order interests, and is immediately followed by a discussion of the primary goods. This suggests that by "the principles of justice" Rawls means either (a) "principles of justice" simpliciter or perhaps (b) "whatever principles the parties choose." But neither alternative yields an unproblematic interpretation of the passage. (a) would be consistent with Rawls' allegiance to the method of pure procedural justice, by stipulating just principles to be whatever is the outcome of the OP choice procedure. But if he is right in arguing that classical and average utilitarian principles would not survive the strains of commitment, and *a fortiori* that only the 2PJ could be the outcome of the OP choice procedure, then the choice of utilitarian principles would not realize this "capacity for an effective sense of justice," and *a fortiori* only the 2PJ could. (b) would be consistent with the possibility suggested above, that the parties might settle on a No-Agreement Point, or choose unjust principles. In this case, again, the parties' capacity for an effective sense of justice would presumably remain dormant. Hence in either case, the parties' a priori motivation to understand, apply, and act from the 2PJ must be built in at the outset.

I am grateful to an anonymous referee for *Social Theory and Practice* for bringing this issue to my attention.

[16] This possibility is supported by the stipulations that (i) the parties are mutually disinterested (TJ 13, 127); and (ii) they are not bound by prior moral ties (TJ 128). If these things are true of them, then why should they commit themselves to choosing principles of justice when each could stand to benefit from injustice? Clearly the additional stipulation of risk aversion is inadequate to answer this question, since each might value the benefits of injustice highly *enough* to risk being victimized by it.

[17] "Instrumentalism, Objectivity, and Moral Justification," *American Philosophical Quarterly* 23 (1986): 373-81.

Is there any remaining justificatory force in the OP for the 2PJ, now that Rawls has made this recent modification? Yes, but this force does not derive from the demonstrated instrumental rationality of the 2PJ; for it is vacuously true that an agent will choose what she has special motivation to choose, other things equal. This fact cannot justify that choice to us unless we, too, have that special motivation, or would have it under appropriate circumstances. Now the relevant circumstances on which the significance and justificatory force of the OP originally relied were supposed to be the practically compelling circumstances of justice, in which the parties were realistically portrayed as competing on an equal basis for scarce resources for achieving their ends, not a highest-order interest in their sense of justice. And it is controversial whether this moral motivation is as essentially reflected in our actual lives; in other words, whether it is in fact more important to us to obtain for ourselves an equitable allocation of the scarce resources we need to survive comfortably, or to develop and exercise our sense of justice. This moral motivation may be reflected more accurately in our idealized self-conception than in our actual motivational structure.[18]

If the 2PJ are to be generated by a highest-order interest that we do not actually have, or have only to a relative minor degree, then the extent of their persuasive force for us will be inversely proportional to the extent of the remoteness of the ideal self-conception that includes that interest from our actual emotions and dispositions. In this case, the justification of the 2PJ will require that we be convinced, first, that the parties in the OP represent characteristics and dispositions that we have, or do or should aspire to have; and second, that the WOS, structured by the 2PJ, depicts the kind of society we envision as a felt social ideal. That is, the fulcrum of Rawls' justificatory strategy must shift from the choice of the parties in the OP to the receptiveness of the reader to the values the OP and the WOS embody. Let us call this the *audience response* conception of justification.

Rawls has developed the audience response conception of justification in recent writings, by explicitly addressing as his audience those who affirm the liberal-democratic tradition his theory of justice also affirms. Clearly, he means to be addressing us as readers about issues of central importance to us, and articulating the basic conditions under which public agreement among us can be achieved:

> [T]his conception of justice provides a publicly recognized point of view from which all citizens can examine before one another whether or not their political and social institutions are just. It enables them to do this by

[18] I develop this distinction at greater length in "Two Conceptions of the Self," *Philosophical Studies* 48 (1985): 173-97; reprinted in *The Philosopher's Annual VIII* (1985).

> citing what are recognized among them as valid and sufficient reasons singled out by that conception itself... [J]ustification is not regarded simply as valid argument from listed premises, even should these premises be true. Rather, justification is addressed to others who disagree with us, and therefore it must always proceed from some consensus, that is, from premises that we and others publicly recognize as true; or better, publicly recognize as acceptable to us for the purposes of establishing a working agreement on the fundamental questions of political justice[19]… Kantian constructivism recasts ideas from the tradition of the social contract to achieve a practicable conception of objectivity and justification founded on public agreement in judgment on due reflection. The aim is free agreement, reconciliation through public reason.[20]

In these passages, the 2PJ are justified, not by their instrumental role in promoting the parties' individual conceptions of the good in the WOS, but rather by their consensual role in providing shared assumptions from the perspective of which we, like the citizens of the WOS, can examine and criticize our social and political institutions. That is, the 2PJ are justified by their function for and connection to us, not the parties in the OP. Hence it is no longer necessary to rely on the continuity of the parties in the OP with the citizens of the WOS for any such justification. Since the 2PJ are no longer conceived as instrumentally rational for the realization of the parties' individual conceptions of the good in the WOS, this more recent conception of justification does not require the truth of the continuity thesis. So we must now turn to the question of whether there are sufficient resources in Rawls' theory of justice to replace it.

III

There is some textual evidence in *A Theory of Justice* that undermines the continuity thesis. The falsity of the continuity thesis is, for example, suggested by Rawls' emphatically drawn distinction between the parties of the OP as "theoretically-defined individuals" and the actual propensities of people in everyday life (*TJ* 147); he asserts that the mutual disinterest in one another of the parties is not necessarily continuous with the motives of "persons in everyday life who accept the principles that would be chosen and who have the corresponding sense of justice" (*TJ* 148). One might speculate that by "everyday life" Rawls may mean here not only the everyday life of the reader,

[19] "Justice as Fairness: Political not Metaphysical," p. 229.

[20] "Justice as Fairness," p. 230.

but perhaps also the everyday life of a person in the WOS. In any case, the latter point is made more strongly in a later discussion of this issue:

> [T]he motivations of persons in a WOS is not determined directly by the motives of the parties in the original position. These motives affect those of persons in a WOS only indirectly: that is, via their effects on the choice of principles. It is these principles, together with the laws of psychology (as these work under the conditions of just institutions), that determine the resulting motivation.[21]

These passages taken conjointly do not, of course, explicitly conflict with all the clauses of the continuity thesis. They are primarily aimed at confuting the purported continuity of the OP's individualistic motivational assumptions with the underlying psychology of members of the WOS. But in so doing the point is also made that the parties in the OP are psychologically discontinuous with the members of the WOS. So whatever prior conceptions of the good, tastes, and interests the parties may have held, their subsequent psychological proclivities nevertheless conform to the constraints of the two principles of justice. Thus we can read these passages as conflicting with clauses (2) and (3) of the continuity thesis.

Clause (1) is called into question by Rawls' claim that "We use the characterization of the persons in the OP to single out the kinds of beings to whom the principles chosen apply" (*TJ* 505; emphasis added). If we understand this to mean that the parties in the OP are not among those particular individuals to whom the principles of justice might apply, but rather schematic adumbrations of the type of individual for whom they are intended -- moral persons -- there is no reason to suppose that the parties bear anything like the concrete relation to particular members of the WOS suggested by the passages adduced in Part 1. On the present construal, the parties in the OP are as much hypothetical constructs to the members of the WOS as they are to the reader; and indeed there is some further evidence to support this possibility.

Near the beginning of *A Theory of Justice*, Rawls relates his conception of the OP to the hypothetical state of nature characteristic of traditional contract theory. After briefly limning what will later become the four-stage sequence in which a conception of justice, then a constitution, and a legislature to enact laws is chosen, Rawls then argues that

> *Our* social situation is just if it is such that by this sequence of *hypothetical agreements* we would have contracted into the general system of rules which defines it. Moreover,... it will then be true that whenever social institutions satisfy these principles, those engaged in them can say to one

[21] "Fairness to Goodness," *The Philosophical Review* 84 (1975): 543.

> another that they are cooperating on terms to which they *would* agree if they were free and equal persons whose relations with respect to one another were fair... the general recognition of this fact would provide the basis for a public acceptance of the corresponding principles of justice. *No* society can, of course, be a scheme of cooperation which men enter voluntarily in a literal sense;... Yet a society satisfying the principles of justice as fairness *comes as close as a society can* to being a voluntary scheme, for it meets the principles which free and equal persons *would* assent to under circumstances that are fair (*TJ* 13; emphasis added).

The hypotheticality of the OP with respect to the vantage point of the reader is reaffirmed in a number of places in *A Theory of Justice* (16, 21, 48, 115, 120, 167, 587). The importance of the above passage lies instead in the facts that first, the continuity thesis is vigorously rejected in its entirety; and second, the implied relation between the parties in the OP and the members of the WOS is very different here from what it is elsewhere. Here the implication is not that the parties afterwards "return to their place in society and henceforth judge their claims on the social system by these principles" (*TJ* 196) with the newly requisite attitudes. Rather, it is suggested that the OP is equally hypothetical relative to any society, including the WOS; and that the former functions as an idealized standard in comparison with which the underlying principles, constitution, legislature, and laws of such a society can be appraised, criticized, and, in theory, revised. On this construal, we regard the WOS not as an immediate and concrete temporal consequence of the series of decisions made throughout the four-stage sequence of the OP and then miraculously implemented by the subsequent efforts of the parties. On the contrary, we view the WOS as a theoretically projected outcome (remote though it may be) of our concerted application of a device, that is, the hypothetical OP and the principles chosen there. This outcome functions as a substantive moral standard for evaluating the circumstances of moral and political conflict that regularly confront us. The OP device both ensures the impartiality of our moral judgments and also yields substantive moral principles in accordance with which we can judge these issues.[22]

Similarly, in the *Dewey Lectures*, Rawls frequently refers to the OP as a construction (*DL* fn. 41, 532), a point of view (*DL* 554, 560, 567), and a framework of deliberation (*DL* 533, 560, 561); and the parties themselves as "agents of construction" (*DL* 547, 552, 560). These characterizations reinforce

[22] Something like this strategy seems to lie behind Rawls' treatment of substantive questions in the second part of *A Theory of Justice*, for example, civil disobedience (Secs. 55-59), but this is too large a topic to discuss here.

the interpretation of the OP as a theoretical construct that enables us as well as citizens in the WOS to perform systematic moral deliberation.

Finally, on page 234 of "Justice as Fairness: Political not Metaphysical," Rawls characterizes his conception of the parties in the OP as "a basic intuitive idea assumed to be implicit in the public culture of a democratic society"; this is part of the "publicly recognized point of view" that enables us to evaluate the justice of our own institutions. He also describes it emphatically as a "device of representation" (236-38) within which

> the conception of justice the parties would adopt identifies the conception we regard -- *here and now* -- as fair and supported by the best reasons... As a device of representation the idea of the original position serves as a unifying idea by which our considered convictions at all levels of generality are brought to bear on one another so as to achieve greater mutual agreement and self-understanding (238; emphasis in text).

In these passages the primary function of the OP is to clarify and articulate our beliefs as citizens participating in the democratic process of institutional evaluation. Similarly, recall *TJ* 13 just cited, which treats the WOS as embodying substantive moral standards against which our actual political institutions could be judged. These two sets of passages, in conjunction with the hypotheticality of the OP relative to the WOS, also suggested by *TJ* 13, effect the almost complete detachment of the parties in the OP as a device we use from the citizens of the ideal WOS generated by it: each bears a continuity relationship, not to the other, but to us.

Thus these passages suggest an alternative to the continuity thesis in the form of the following conditions:

> (i) The parties in the OP do not regard themselves as future members of the WOS (clause (1) and (3) of the continuity thesis). Rather, they recognize themselves to be psychologically discontinuous with those members, and recognize also that their choice of principles determines the general motivational features of the members of the WOS.
>
> (ii) The OP itself is not an actual event (clause (2) of the continuity thesis), but rather a hypothetical one relative to the WOS. Hence the parties are not in fact physically continuous with any member of the WOS.

Let us call this the *discontinuity thesis*.

A number of exegetical consequences follow from adopting the discontinuity thesis as the favored interpretation of the OP. If we assume the falsity or irrelevance of the continuity thesis to the Rawlsian enterprise, those criticisms that presuppose it must be disregarded. One cannot then argue against the conception of the OP or the two principles of justice chosen there on the grounds of what the parties so situated would do or think after they

got out of it or before they went into it,[23] or how their society might be colored by their prior psychological proclivities. Without the continuity thesis, there is no *prima facie* reason for the parties in the OP to suppose that anyone's needs or conception of the good might conflict with, undermine, or be frustrated by the constraints imposed by the two principles of justice on the basic structure of society. This is not to propose that everyone's psychology in the WOS must be consonant with these principles. It is just to argue that without the notion of continuing persons, this possibility does not suffice to deter the parties in the OP from choosing as they are presumed to do. It does not suffice because the parties then have no reason in deliberating to provide for the possibility that the persons they turn out to be might be persons to whom their choice in the OP is unacceptable. If the continuity thesis fails, the parties in the OP must be regarded as *self-determining* in the very strong and liberal sense that in the OP, they determine the kinds of persons they are to be, the kind of psychology they will have, and the kinds of moral constraints they will be prepared to accept, by deciding what principles of justice are to regulate their interactions. The circumstances of the OP must then be viewed as a radical discontinuity in their adult lives, after which they become the kinds of persons who are constrained and partially determined, not by the continuity of their previous psychological histories, but by their choice of moral principles in the OP.

At the same time, those passages we have cited from *A Theory of Justice* and *The Dewey Lectures* that lend support to the continuity thesis must be similarly bracketed. We must, for example, interpret Rawls' discussions of the expectations of the parties in the OP in the same hypothetical light as we do the concept of the OP itself. The parties must be conceived, and must conceive themselves, as deciding on principles *as though* such principles were to govern their life prospects. For they recognize that they are in fact choosing principles not for themselves, properly speaking, but for the persons they thereby choose to become. Thus they must regard themselves as advancing in the subsequent society not only their conceptions of the good, but indeed their idealized conceptions of the self which the choice of principles determines.

Rawls' arguments regarding the strains of commitment must be qualified in much the same way: the issue then becomes not whether the parties can adhere to the chosen principles, but instead whether the preferred conception of the self includes this capacity. This implies, first, that the capacity for a

[23] Certain of Ronald Dworkin's arguments in "The Original Position" (*University of Chicago Law Review* 40 [1973]: 500-33; reprinted in Daniels, pp. 16-52) against the justificatory function of the OP that depend on his distinction between "antecedent" and "actual" interest (pp. 20-21) would necessitate revision on this interpretation.

sense of justice cannot be stipulated as a motivational assumption of the OP, independently of this preferred conception. Second, it implies that the capacity for a sense of justice cannot be used as a criterion for differentiating between acceptable and unacceptable principles of justice. For we can expect a great variety of such principles to be successful in tailoring a conception of the self that will stably adhere to them.

Finally, the abandonment of the continuity thesis entails the abandonment of the instrumentalist strategy of justification that is, for many, centrally definitive of the contract-theoretic tradition. That tradition is founded on the reasoning that a justified society is one by the rules of which individuals who are instrumentally rational and self-interestedly motivated to improve their lot in the state of nature would agree to be bound, in order to regulate their interactions. In this picture, the state of nature, the self-interested and instrumentally rational individuals, and the social contract are jointly continuous but hypothetical relative to our actual society. But by abandoning the continuity thesis, the modifications we have traced in Rawls' recent writings functionally eliminate the state of nature, the self-interest, and the instrumental rationality of the individuals; and stipulate the hypotheticality of the contractual agreement in the OP relative to the ideal WOS itself. There are many elements in Rawls' thought that continue to affiliate him with the tradition of social contract theory. But his evolving conception of justification represents a departure and an innovation relative to that tradition.

IV

In closing, I want to consider the implications of the discontinuity thesis, first, for Rawls' view on personal identity; and second, for his concept of wide reflective equilibrium. A recent criticism of Rawls' theory has focused on the seeming disparity between claims made in *A Theory of Justice* supporting the choice of the 2PJ over utilitarianism, and Rawls' more recent treatment of the issue of the relevance of the problem of personal identity to moral theory.[24] In *A Theory of Justice*, it was suggested that the fact that utilitarianism does not take seriously the distinction between persons might be a reason why the parties in the OP would be disinclined to choose it. For they would then have good reason to doubt whether a society erected on utilitarian first principles would protect and promote those long-term plans and interests which the parties each know themselves to have (*TJ* 27-29). In "The Independence of

[24] Samuel Scheffler, "Moral Independence and the Original Position," *Philosophical Studies* 35 (1979): 397-403.

Moral Theory," on the other hand, Rawls is concerned to show that conclusions in the philosophy of mind concerning personal identity, that is, that it involves bodily continuity and also mental continuity subject to varying degrees of fluctuation, do not conclusively favor one moral theory over any other. Thus Rawls claims that

> what sorts of persons we are is shaped by how we think of ourselves and this in turn is influenced by the social forms we live under... There is no degree of connectedness that is natural or fixed; the actual continuities and sense of purpose in people's lives is relative to the socially achieved moral conception.[25]

But Scheffler has tried to show that this then implies either that (1) the parties cannot assume themselves to have long-term plans and purposes -- since this feature would characterize a particular kind of personal identity which is no more natural or fixed than a weaker one in which plans, projects, memories, and experiences undergo continual alteration and replacement, and hence they cannot choose principles of justice with an eye to protecting such long-term plans and purposes; or else (2) the parties' choice of the 2PJ on the supposition of having long-term interests demonstrates only that individuals having a certain kind of personal identity would choose a society that would protect it, without providing any independent argument against the choice of utilitarianism.[26] If the parties are then not assumed to have long-term plans and purposes, but nevertheless are assumed to choose principles of justice for the basic structure of the society in which they will live, it is then an open question whether they would choose a society which protected long-term interests over one that did not.

But if the continuity thesis is supplanted by the discontinuity thesis, and in particular clause (i), these difficulties do not arise. For it is only if the parties are conceived as continuing persons who had adopted certain projects and purposes prior to the OP, which they then advance in the WOS subsequent to it, that there is any independent requirement for how long a person in the OP must endure, and how long a long-term interest must be, in order to count as long-term. A person, and hence her goals and interests, must endure long enough to have originated and engendered in the person a deep commitment to the fulfillment of these goals and interests before the circumstances of the OP occurred; they must survive the protracted period of conflict, dialogue and deliberation which the OP, with the support of clause (ii), surely entails; and they must survive the actual lengthy period of

---

[25] Rawls, "The Independence of Moral Theory," p. 20.

[26] Scheffler, "Moral Independence and the Original Poition," 399-401.

implementation of the 2PJ in the WOS which the continuity thesis plus the four-stage sequence entails. Such longstanding commitment to a goal or interest is impressive indeed. It might even survive what we normally think of as a natural human lifespan.

Rejecting the continuity thesis, on the other hand, permits us to leave open the question of how long, in actual time, the parties' long-term interests must be in order to count as long-term. It is then sufficient that a long-term interest survive for the duration of a person's adult life, as we would normally expect. But there is now no reason to place any prior constraints on how long such a life must be. Hence whether a person has a weak identity or a strong one is irrelevant to whether that person can be said to have long-term interests or not. The person's interests are identified as long-term relative to the duration of her personal identity. Now since there are no longer any independent constraints on how long the parties themselves endure, nothing about the conception of the OP forces the characterization of the parties as having particularly weak or particularly strong personal identities. And the ascription to them of long-term interests fails to decide this question one way or the other.

Furthermore, that a person has a weak personal identity in any case does not imply a lack of concern with personal survival. Even if I know that my character is so volatile and unstable that I can realistically expect to be a completely different person in five years, I need not be happy about this. Or I may wish my interests to endure for as long as I do, however long that is. It is both conceivable and likely that an individual with a weak personal identity would be either concerned with her own person survival, or hold personal survival as a value in general, or both. And so long as the continuity thesis is rejected, the question of how long, in real time, such an individual would consider it in general valuable for a person to survive, is once again left open. So the fact that the parties in the OP may have weak personal identities does not imply that they would have no reason to choose principles of justice which would respect and protect long-term interests. Indeed we might expect the concern with personal survival to increase with the threat to personal survival. Thus that utilitarianism fails to take seriously the distinction between persons and hence would fail to protect their interests remains a good reason for any party in the OP not to choose it.

Now clause (i) of the discontinuity thesis implies that the parties know they will not psychologically survive past the circumstances of the OP. They regard themselves as determining future selves for the WOS in light of the chosen principles of justice, and hence as determining the long-term goals and interests these selves will have. Hence they choose principles, not with an eye to promoting instrumentally their own long-term interests, but rather with an eye to protecting the long-term interests of the kind of person they

simultaneously choose to become. And to suppose that the parties could be unconcerned about the survival of those long-term interests -- however long in real time they might be -- would be to suppose that they were indifferent to the particular nature of the self they had chosen. But since they regard themselves as self-determining, there is no reason to believe they would be.

Thus Rawls' claim, that conclusions about the nature of personal identity are irrelevant for the construction of a moral theory, can be made to hold in spite of the apparent conflict with the earlier argument against utilitarianism. These conclusions are irrelevant as long as the characterization of the OP is not biased by the assumption of the continuity thesis. For only then is Rawls committed to a type of personal identity which the parties might desire self-interestedly to prolong into the WOS, rather than one they desire disinterestedly to create.[27]

But perhaps the most important consequence of supplanting the continuity thesis is that greater attention needs to be directed toward Rawls' concept of wide reflective equilibrium (henceforth WRE). In *A Theory of Justice*, the concept of WRE referred to a stable state in which the description of the OP and the principles chosen in it to govern the WOS had been mutually adjusted and compared with other alternatives so as to finally match our considered moral judgments (*TJ* 20, 4849). As we have seen, many of the features originally ascribed to the OP require modification in order to circumvent the unacceptable implications of the continuity thesis. In particular, the motivational features of the OP that, at least on the interpretation offered in Section 1 of this discussion, lead the parties to choose principles of justice in order to advance their conceptions of the good in the subsequent WOS must be abandoned, and replaced by some other connection between the OP and the WOS. Now clause (ii) of the discontinuity thesis denies that the major connection between the OP and the WOS is mediated by continuing persons who are assumed to participate in both, and we have already seen that Rawls himself has begun to give increasing prominence to us, the audience, conceived as members of a liberal democratic society who are reflective, self-critical, and morally concerned thinkers who attempt to give coherence, substance, and reality to their considered moral judgments. As moral mediators between the OP and the WOS, we advert to the OP in order to attain the requisite impartiality of judgment, and to the WOS in order to substantiate and specify the scope of application of those judgments

[27] However, my conclusions here should not be taken to endorse Rawls' recent disavowal of the relevance of metaphysical questions to his theory of justice *tout à fait* ("Justice as Fairness: Political not Metaphysical," pp. 230, 238-40, fn. 22). If my arguments are well-taken, Rawls' disavowal is too sweeping.

themselves. Hence the attainment of WRE must be measured by the internal coherence of our own moral judgments on the hand, and the point of view from which we make them on the other. The importance of this line of thought to Rawls' thinking from the very beginning is strongly suggested by his closing remarks in *A Theory of Justice*:[28]

> Once we grasp this conception, we can at any time look at the social world from the required point of view... This perspective of eternity is a certain form of thought and feeling that rational persons can adopt within the world... and arrive together at regulative principles that can be affirmed by everyone as he lives by them, each from his own standpoint. Purity of heart, if one could attain it, would be to see clearly and to act with grace and self-command from this point of view (*TJ* 587).

[28] For this reason I now think I was too hasty in claiming (in "A Distinction Without a Difference," *Midwest Studies in Philosophy VII: Social and Political Philosophy* (Minneapolis, Minn: University of Minnesota Press, 1982), p. 406) that Rawls' theory of justice contains no action-guiding part at all. There has always been a practical and applied strain in his thought which is becoming increasingly salient in his more recent writings.