The Original Position from <u>Theory</u> to Political Liberalism

> "[I]t is impossible to finish [Rawls's] book without a new and inspiring vision of what a moral theory may attempt to do and unite; of how *beautiful* a whole theory can be."
>
> -- Robert Nozick, *Anarchy State and Utopia*, p. 183

In his introduction to the first edition of *Political Liberalism*, Rawls said that much of the criticism of *A Theory of Justice* sprung from a failure to see that the original position had been used in that book as a "device of representation". (*PL*, p. xxxvi) The volume of critical attention that has quite understandably been lavished on Rawls's use of the original position to defend his two principles of justice can make us forget how versatile a device the original position was. Indeed the wide range of, and underlying connections among, the arguments in which Rawls appealed to the original position – and the beauty of the result – all contribute to making the account of justice laid out in *TJ* a *theory* properly so called.

Some of the connections Rawls asserted among seemingly unrelated problems are connections he drew in order to show that the well-ordered society of justice as fairness would be stably just. He sometimes expressed his desired conclusion by saying that the principles would "generate their own support". I shall suggest, however, that it is illuminating to think of stability as brought about, not by the *principles of justice*, but by *the agreement on them*. More precisely, I shall suggest that it is illuminating to think of the agreement reached in the original position as a special kind of what is sometimes called a "self-enforcing agreement".[1] One of the things that is illuminated by thinking of it this way are Rawls's reasons for connecting the diverse parts of his theory *by means of the original position*.

Rawls's reasons for doing this still need to be brought fully to light more than four decades after *TJ* was first published. For in *TJ*, Rawls had observed that he could have dispensed with the original position and argued directly for the principles of justice from the conditions that underlies his social contract. (*TJ*, p. 138) The question of whether the original position is an essential part of Rawls's theory, and in what ways it

[1] Speaking of his own theory, Rawls says "the psychological construction as a whole is self-reinforcing". The remark occurs at his *CP*, p. 106.

is a helpful one, have interested scholars ever since.[2] Even the best answers to these questions stress the ways in which appeal to the original position makes Rawls's arguments for the two principles clearer and more economical than they would otherwise be. But in doing so, they fail to mention other arguments from which it is less clear that the original position could be eliminated -- arguments in which Rawls, I shall contend, uses the representative function of the original position to show how an agreement on the principles would stabilize or enforce itself.

To substantiate my claim that Rawls did in fact argue for this conclusion, I shall begin by reviewing what Rawls meant by referring to the original position as a device of representation. I then look at the conditions of a self-enforcing agreement and say how those conditions need to be strengthened for the special case of an agreement on principles of justice. I show how Rawls used the original position, in effect, to show that the agreement reached in it could satisfy those conditions. I also contend that Rawls was able to use the original position in the arguments because of what he devised it to represent. As is well known, Rawls says he made the turn to political liberalism because he came to think that the stability arguments of *TJ* failed. I shall explain that failure as the failure of the agreement reached in the original position to satisfy two of the conditions of self-enforcement. That failure necessitated a change in what the original position represented. This change opened the possibility of a different argument for stability than the one Rawls had offered in *TJ*, an argument which appealed to the possibility of a very different self-enforcement mechanism.

First, however, a few words about the texts on which I shall focus. My topic is not the original position in political liberalism generally, but the original position in *Rawls's* political liberalism. I take that to be, not the original position in Rawls's book *Political Liberalism*, but the original position in Rawls's conception of justice, justice as fairness, when presented as one instance of political liberalism. Since Rawls began presenting justice as fairness as a political liberalism in "Justice as Fairness: Political not Metaphysical", I shall draw on his writings from that essay onward. Rawls's use of the original position before his political turn is introduced at length for purposes of contrast. In drawing that contrast I shall confine myself largely to *TJ* and the original Deweys, leaving aside the use and development of the original

[2] See Dworkin; Joshua Cohen, "Democratic Equality".

position in Rawls's earlier essays. Moreover, since the 1999 edition of *TJ* incorporates revisions which reflect developments in Rawls's thought after the book's initial publication -- including revisions to arguments in which the original position is used -- I shall treat the 1971 edition as the definitive statement of justice as fairness in its pre-political form. Throughout I shall ignore the role of the original position in Kantian constructivism, which requires a separate treatment from the questions I shall take up here.

§1. The OP as a Device of Representation

When Rawls said that much of the criticism of *TJ* sprung from misunderstanding his use of the original position, the criticism he had in mind was that *TJ* had relied upon "an abstract conception of the person and use[d] an individualist, nonsocial, idea of human nature". (p. xxxvi) To see just what he meant by claiming that he had used the original position as a device of representation in *TJ* -- and to see how he later used it as one in political liberalism -- it will help to see how the claim constitutes a response to that criticism. To see *that* it will help to spell out the criticism in its bluntest form.

Let us start with what Rawls calls an "idea of human nature". And let us take the contents of such an idea to be the essential properties of human beings, understood as the properties which all and only human beings have in all times, places and possible worlds. An idea of human nature which has these contents would seem to be metaphysical in at least this sense: the idea would seem to have implications for questions which are traditionally identified as metaphysical, such as questions about as the identity conditions of human beings over time and across worlds. The criticism that Rawls relied on such an idea does not hold that Rawls explicitly argued for one in *TJ* or that one falls out of his work even if he did not argue for it. Rather the criticism alleges that crucial arguments in *TJ* presuppose a metaphysical idea of our nature that is objectionably individualist. More specifically, the objection alleges that in using the original position, Rawls was tacitly relying on such an idea.

The basics of Rawls's response to this objection are well understood. Rawls chose the characteristics with which he endowed the parties in the original position – their rationality, mutual disinterest, desire for primary goods and interest in the subsequent development and exercise of their moral powers – and imposed a veil of ignorance on them, so that the original position would be "a state of affairs in

which the parties are equally represented as moral persons". (*TJ*, p. 122) And so in designing the original position, and in using it in the arguments of *TJ*, Rawls did rely on and elaborate a basic intuitive idea of the human. But the contents of that idea are not properties each of which is metaphysically necessary for counting as a human being. Rather, the idea was what we might call a basic intuitive "idea of moral personality".[3]

In *TJ*, the original position functioned as a device for representing the contents of that idea. More precisely, it functioned as a device for representing human beings as possessed of the defining features of moral personality – for short: it functioned as a device for representing us as free and equal persons. The original position functioned that way because Rawls devised it for precisely that purpose. And he devised it for that purpose because he wanted to identify principles of justice for the basic structure of a liberal democracy, and he thought that the way to identify them was to identify principles that would be chosen by moral persons fairly situated. And so Rawls framed the idea of moral personality which underlies the original position to solve problems in the theory of justice, not a set of problems in metaphysics. Rawls's reasons for framing the idea of moral personality do not rule out the possibility that that idea is also a metaphysical idea of human nature. But they do show that he did not put it forward as such an idea and that it does not need to be one to serve his philosophical purposes.

§2. What Makes a Theory?

Rawls's philosopihcal purposes are extremely ambitious, for his overarching aim was to develop a theory of distributive justice. To appreciate the magnitude of that ambition, it will help to recall what qualifies Rawls's account as a theory – indeed, what qualifies it as an example of what Quentin Skinner called "Grand Theory".[4]

---

[3] In "Political not Metaphysical" Rawls said that justice as fairness is worked up from basic intuitive ideas found in the public culture, and it is only after that essay that the phrase assumed prominence in his work. But even in *TJ*, Rawls had said "I have tried to classify and to discuss conceptions of justice by reference to their basic intuitive ideas, since these disclose the main differences between them," (*TJ*, p. 52), though he did not identify the source from which those ideas were drawn.

[4] *The Return of Grand Theory in the Human Sciences* (Cambridge University Press, 1985) ed. Quentin Skinner; the chapter on Rawls, by Alan Ryan, is found at pp. 101-20.

Rawls relied on the abstraction and idealization characteristic of theories – such as the abstraction from illness and physical infirmity and the idealization of perfect compliance -- to distinguish the questions he considered central from those he considered secondary or peripheral, and to render the problems that interested him more tractable. Rawls's account is also grandly theoretical in its scope, the breadth of which is shown by the three parts of *TJ*. Thus in Part I Rawls develops an elaborate argument for his own principles of justice and against utilitarianism, in Part II he considers at some length how those principles can be institutionalized, and in Part III he argues that the principles would be stabilized by citizens' sense of justice and by the congruence of the right and the good. Another feature of the account presented in *TJ* that qualifies it a grand theory is its highly articulated structure, which one early reviewer likened to that of a gothic cathedral.[5] For Rawls moved through the three parts of *TJ* systematically, identifying the main problems in which he was interested and then posing and addressing subsidiary questions those problems raised. He also proceeded cumulatively, building later arguments on the conclusions of earlier ones.

Finally, there is a methodological unity to Rawls's account of justice. Students and casual readers of *TJ* may think that the original position was used only to defend the principles of justice. In fact, Rawls used the original position to attack problems in all three of the account's main parts. Once brought to light, this methodological unity raises the question of whether the problems and topics Rawls used the original position to attack have some underlying connection in virtue of which he thought they all would yield to the same approach.

It might seem that they do not. Rawls himself observed in *Political Liberalism* that "the problem of stability has played very little role in the history of moral philosophy" (*PL*, p. xix), thereby seeming to concede that philosophers have generally thought the problem he took up in Part III of *TJ* is unrelated to those he took up in Part I. He also allowed that the problems of institutional design taken up in Part II might not seem to be *philosophical* problems at all. (*TJ*, p. 95) But history and appearances notwithstanding, Rawls asserted a deep connection among the problems taken up in the three parts of *TJ*. His elegant

---

[5] John Chapman, "Rawls's Theory of Justice", *American Political Science Review* 69,2 (1975): 588-93, p. 588.

unification of those problems, reflected in the methodological unity of his attempts to answer them, gives the account of *TJ* the aesthetic appeal of a grand theory. The appeal and, I shall contend, the power of the theory would be dissipated if the original position were dispensed with.

§3. What Unifies the Theory?

What is the underlying connection among the problems Rawls took up in the three sections of *TJ*?

I said at the outset that it is illuminating to read Rawls as arguing that the agreement reached in the original position would be a "self-enforcing agreement". The notion of self-enforcement is at home in the theory of non-cooperative games. These are games in which players can communicate, but in which agreements reached on the basis of such communications are not binding because there is no mechanism – such as an agent with coercive power over the players -- for enforcing any agreements they may reach.[6] Crudely put, an agreement reached under these conditions is *self-enforcing* when parties to the agreement freely uphold it, and when the agreement itself is what insures that its terms are freely upheld.

Somewhat more precisely, there are three conditions a self-enforcing agreement must satisfy. First, trivially:

(1) The principles that the players are to follow – the strategies they are to play – must be specified by the terms of an agreement or contract among them.

Second, if the agreement is actually to be upheld, then none of the parties to it can deviate from its terms. And since deviation is not prevented by an enforcement mechanism, the terms agreed to must be an equilibrium. And so:

[6] The locus classicus of the distinction between cooperative and non-cooperative games is John Harsanyi, "A General Theory of Rational Behavior in Game Situations", *Econometrica* 34,3 (1966): 613-34, p. 616.

(2) None of the parties to the contract referred to in (1) can have sufficient reason to deviate from its terms, at least so long as all the other parties comply with them.[7]

But this condition is too weak. Games can have multiple equilibria, and an agreement on terms which define one of those equilibria will succeed in enforcing itself only if it is an agreement on terms that parties to the agreement actually honor. Indeed, the point of making an agreement might be to coordinate on the selection of one equilibrium-state among many. This suggests a strengthened version of (2) according to which:

(2') None of the parties to the contract referred to in (1) can have sufficient reason to deviate from its terms, at least so long as all the other parties comply with them, and all the others do comply.

Third, an agreement to do what parties are going to do anyway is otiose, and is therefore not an agreement which brings about or enforces compliance with its own terms.[8] And so:

(3) The fact that the terms were agreed to in the contract referred to in (1) must be what brings it about that all the parties comply with them.

---

[7] This statement of the second condition does not require that no coalition has sufficient reason to deviate from the agreement. Coalition-proof equilibria have been extensively studied in recent decades, and a number of equilibrium concepts have been proposed and defended. In stating the second condition, I gloss over the question of which is the most reasonable concept to use in defining self-enforcing agreements. I do so because the conditions on self-enforcement are stated to introduce the stronger conditions of what I shall call "Rawlsian self-enforcement". While I believe Rawls's arguments could be extended to show that agreements which are Rawlsian self-enforcing are coalition-proof, I also believe that the possibility of individual deviation was the one he had most clearly in view and that extending his arguments to cover deviation by groups would take us too far afield.

For a classic treatment of coalition-proof equilibria, see B. Douglas Berheim, Bezalel Peleg and Michael D. Whinston, "Coalition-Proof Nash Equilibria: I. Concepts", *Journal of Economic Theory* 42 (1987): 1-12. For an illuminating exploration in connection with Locke's social contract, see Joshua Cohen, "Structure, Choice and Legitimacy", *Philosophy and Public Affairs* 15,4 (1986): 301-24.

[8] Here I rely on Robert Aumann, "Nash Equilibria are not Self-Enforcing", in Aumann, *Collected Papers: Volume I* (Cambridge, MA: MIT Press, 2000), pp. 615-20.

The agreement which Rawls wanted to show would be self-enforcing is an agreement on principles of justice. These are moral principles, needed to play a public role and capable of moving us to act for their own sake. The three conditions on self-enforcing agreements need to be strengthened for that special case.

First, Rawls thought that an agreement on principles of justice had to be reached under the special conditions. What he calls "the guiding idea" of his theory says what those conditions are: according to that "guiding idea," he says, principles of justice "are the principles that free and rational persons concerned to further their own interests would accept in an initial position of equality." (*TJ*, p. 11) And so while (1) is a trivial condition on self-enforcing agreements, the strengthened version of it is the decidedly non-trivial:

(R.1) The principles of justice that members of the well-ordered society are to follow must be specified by the terms of an agreement or contract among them, conceived as "moral persons, as creatures having a conception of their good and capable of a sense of justice." (*TJ*, p. 19)

To see how the second condition should be strengthened, it helps to recall what is attractive about self-enforcing agreements. Self-enforcing agreements are attractive because they are maintained without coercion, and so are freely honored in some sense of 'freely'. The agreement reached in the original position is to be honored by members of a society whose basic institutions satisfy the principles. Rawls wants them to maintain the agreement while exercising a kind of freedom which is especially robust, freedom which characterizes the lives of moral persons as such. The strengthened version of (2′) has to require that the agreement be honored by conduct which realizes that kind of freedom. This suggests that the strengthened version of (2′) should be a conjunction, one conjunct of which expresses the requirement:

(R.2') None of the members of the well-ordered society can have sufficient reason to deviate from the principles of justice, at least so long as all the others comply with them, and all do comply by conducting themselves as "moral persons, as creatures having a conception of their good and capable of a sense of justice." (*TJ*, p. 19)

Finally, the strengthened versions of (1) and (2) suggest a ways of strengthening (3) which yields:

(R.3) The fact that the principles were agreed to in the contract referred to in (R.1) must be what brings it about that members of the well-ordered society comply with it, as (R.2′) requires, and the connection between the contract and the conduct referred to by (R.2) must be established in ways which treat members of the well-ordered society as "moral persons, as creatures having a conception of their good and capable of a sense of justice." (*TJ*, p. 19)

But Rawls insists that the agreement on his principles is and must be hypothetical rather than actual (*PL*, pp. 271ff.) and this poses a serious problem for this way of strengthening (3). To see why, it will help recall the rationale for that condition.

As I said when I introduced (3), self-enforcing agreements cannot be otiose or redundant, as an agreement would be if it was simply a contract to do what the parties were going to do anyway. Crudely put, (3) responds to the intuition that an agreement can be self-enforcing only if the agreement itself makes a difference to what the parties do. (R.3) incorporates this response by requiring an efficacious connection between a self-enforcing agreement and the conduct which maintains it.

There are at least two ways an agreement could make a difference, and hence two ways in which the connection required by (R.3) could be efficacious. First, someone's agreement to execute the terms of a contract can itself provide her sufficient reason to comply, just as a promise can. But this is not the kind of difference that self-enforcing agreements are supposed to make. Rather, self-enforcing agreements are agreements in which the payoffs provide the parties sufficient reason to comply. This brings us to the second way in which agreements can make a difference, by bringing it about that the expected payoffs of the agreement provide such reasons.[9] When agreements make this second kind of difference, they exercise causal power rather than reason-giving force. That *this* is the kind of difference self-enforcing agreements

[9] Aumann's example is one in which a pre-play agreement affects each player's choice of strategy by conveying information about which strategies others intend to play.

are supposed to make is suggested by word ‘enforcing’, which denotes a causal notion. But agreements which are merely hypothetical cannot exercise causal power. Since the connection (R.3) requires is a causal connection, it states a condition which the agreement reached in the original position cannot satisfy. Moreover, because hypothetical agreements cannot exercise causal power, they cannot enforce themselves. And so the claim that the agreement reached in the original position would be self-enforcing seems mistaken, and the attempt to identify conditions under which it would be seems wrong-headed from the start.

Though non-actual agreements cannot exercise causal powers, institutions that satisfy and are known to satisfy the terms of an agreement, and that are modeled on the conditions in which such an agreement would be made, *can* exercise them. One way they can do so is by educating those who live under them so that they come to find the payoffs of honoring the agreement high enough to comply. That education connects the hypothetical agreement and conduct that honors its terms, but there are constraints the connection must satisfy. And so the proper way to strengthen (3) yields not (R.3) but:

> (R.3') The fact that the principles would be agreed to in the contract referred to in (R.1) must be what brings it about that members of the well-ordered society comply with them, as (R.2′) requires, and the connection between the hypothetical agreement and the conduct referred to by (R.2′) must itself be established in ways which treat members of the well-ordered society as “moral persons, as creatures having a conception of their good and capable of a sense of justice.” (*TJ*, p. 19)

Conditions (R.1), (R.2′) and (R.3′) are jointly sufficient for an agreement to be what I shall call “Rawlsian self-enforcing”. In the next three sections, I shall argue that an agreement reached in the original position meets these conditions. That means that Rawls uses the original position, not just to identify the right principles of justice, but also to say how the principles should be institutionalized so that they would be freely complied with. As we shall also see, he was able to use the original position in these ways because it represents us as free and equal moral persons. The fact that the original position could be used in all of these ways reflects the further fact that Rawls asserts a deep connection between justice and stability: the right

principles of justice must be acceptable to us as moral persons, and principles which are acceptable to us are also principles we can be brought freely to honor as such persons.[10] Thus the underlying unity of subject-matter which helps to make Rawls's account a systematic *theory* of justice is revealed by the fact – missed by the critics to whom I referred at the outset -- that Rawls contrived the original position as a device for representing our moral personality.

§4. Satisfying (R.1)

(R.1) requires

> (R.1) The principles of justice that members of the well-ordered society are to follow must be specified by the terms of an agreement or contract among them, conceived as "moral persons, as creatures having a conception of their good and capable of a sense of justice." (*TJ*, p. 19)

Rawls's argument that his two principles of justice would be chosen in the original position has been as extensively as any argument in contemporary philosophy. And so I believe it is well understood by now that Rawls thought the original position is a "philosophically favored interpretation of this initial choice situation for the purpose of a theory of justice" (*TJ*, p. 18) – and, in particular, for the choice of principles. It is therefore well understood that Rawls thought that the agreement reached in the original position satisfies (R.1). Moreover, it is well understood that the original position is philosophically favored, and hence that (R.1) is satisfied, because the original position represents the features of moral personality. Since these points are well understood, I shall not belabor them here.

Rawls makes an assumption which, if it could be vindicated, would lend further support to the conclusions that agreements reached in the original position satisfies (R.1). That is the assumption that the original position is an appropriate initial situation for choosing other principles of right, such as the principles of obligation and of natural duty. I shall take it as clear enough that the assumption *can* be vindicated, at

[10] I explore contractualism's connection between justice and stability in "Relational Equality, Inherent Stability and the Reach of Contractualism", *Social Philosophy and Policy* (forthcoming).

least if the original position can shown to be the appropriate situation for choosing principles of justice. What might not be clear is that the original position is the appropriate for choosing these other principles of right because of what it represents. To suggest that it is, I want to call attention to a crucial premise in that argument which might easily be overlooked, since it is found in a relatively neglected section of *TJ* -- section 23 on the formal constraints on the concept of right. One of those constraints is the universality condition, according to which all principles of right "must hold for everyone in virtue of their being moral persons". (*TJ*, p. 132) This constraint can be satisfied only if such principles are adopted in a choice situation in which those who are to be subject to the principles are represented as moral persons, for if they were represented as having features in addition to or instead of those definitive of moral personality, the principles would hold of them in virtue of those other features and the universality condition would then be violated. Thus what makes the original position an appropriate initial situation for adopting principles which satisfy this condition – hence any principles of right at all -- is its representative function.

While the arguments to which I have just referred may show that the original position is "*a* philosophically favored interpretation of this initial choice situation for the purpose of a theory of justice" (*TJ*, p. 18, emphasis added), they do not show that it is "*the* philosophically favored interpretation" – favored, that is, relative to contract situations that would lead to the choice of other principles of justice and to other contract situations that have been proposed in the history of philosophy. I now want to look at arguments for that conclusion more closely. For these arguments demonstrate the versatility of the original position. They also provide further and less familiar grounds for the conclusions that the agreements reached in the original position satisfy (R.1) and satisfy it because of what the original position represents.

Rawls recognizes that contractualism is a distinctive approach to the identification of principles of justice, and that alternative conceptions such as utilitarianism have identified principles for basic institutions in different ways. Yet Rawls "conjecture[s] that for each traditional conception of justice there exists an interpretation of the initial situation in which its principles are the preferred solution." (*TJ*, p. 121) In *TJ*, sections 28 and 30, Rawls tries to show that that conjecture is true of average and classical utilitarianism. In doing so, he shows that the two conceptions "derive from markedly distinct assumptions". (*TJ*, p. 189)

The "markedly distinct assumptions" from which average and classical utilitarianism are derived are assumptions about how parties to the social contract should be represented. Thus Rawls says that the principle of average utility would be chosen by "a single rational individual (with no aversion to risk)" who entered the initial choice situation and tried to "maximize his own prospects", while the classical principle of utility would be adopted in a contract among "perfect altruists." (*TJ*, p. 189) The ways in which these two versions of utilitarianism think contracting parties should be represented are quite different from the way they are represented in justice as fairness. The three conceptions of justice therefore begin with three different views about which features of moral personality are relevant to theorizing about justice, views which can be brought to light and compared by starting with the original position and asking how it would have to be varied to yield different principles. The original position yields this deeper understanding of the two forms of utilitarianism, and makes it possible to compare the basic intuitive ideas of moral personality from which they are derived with the basic ideas of justice as fairness, because the original position – along with its utilitarian variants – is a device for representing what are alleged to be the features of moral persons.

The original position also enables Rawls to pinpoint and to remedy the shortcomings of competing views which, like justice as fairness, take principles of justice to be the object of agreement. For example, Rawls briefly contrasts the agreement reached in the original position with an equilibrium reached as a result of agreements made among willing traders with full information. (*TJ*, pp. 119-20) He contends that the former agreements but not the latter are just "whatever they turn out to be" because of the conditions of the original position. In "Basic Structure as Subject", he contrasts the social contract reached in the original position with that reached in Locke's state of nature. He argues that the latter unjustly deprives the property-less of the right to vote because the well-off in Locke's state of nature know, and so are in position to use the power that accrues as a result of, their property holdings. (*PL*, p. 287)

The problem with both economic markets and a Lockean state of nature is that they allow the content of agreements reached in them to be affected by differences in bargaining power.[11] This is true even of self-enforcing agreements reached in them, and is precisely why condition (1) on self-enforcing

[11] On the problem with Locke, see Cohen, "Structure, Choice, Legitimacy".

agreements proved too weak for Rawls's purposes. The original position improves on markets and on Locke's state of nature by using the veil of ignorance to block the influence of potential differences in bargaining power. As a result, the agreement reached in the original position satisfies the stronger and more plausible (R.1). The conditions of the original position are appropriate for the choice of principles of justice because they force the content of agreements reached in to depend only upon the shared features of moral personality. As Rawls puts it in a passage to which I shall return in the next section, in the original position our "nature [is] the decisive determining element" of the choice. (*TJ*, p. 253)

Rawls also thinks that the original position enables him to answer what seems like a powerful objection to Kant's ethics. Kant thought that principles of right are those which would be adopted by our noumenal selves. Because our noumenal selves are outside the order of nature, acting from such principles would be acting on principles we truly give ourselves. Henry Sidgwick objected that our noumenal selves are completely unconstrained in their choice of principles, so that acting on any consistent set of principles at all – including the principles of the scoundrel -- would seem to count as autonomous. But, Sidgwick thought, to choose the scoundrel's principles is surely not to choose as a free and equal rational being. The conditions that characterize noumenal selves are therefore not strong enough for the choice principles of right, contrary to what Kant thought. This matters for present purposes because if Sidgwick's objection is sound, an agreement reached under those conditions would seem to violate (R.1).

Rawls does not put the point exactly this way, but I take his response to be the following. Though the scoundrel's principles may have been freely chosen, his actions do not express his autonomy. To live autonomously in the world requires living in certain ways among other rational but vulnerable beings, ways which proceed from a subset of all the principles which might have been freely chosen. Kant left himself open to Sidgwick's objection because he did not have a way of imposing this requirement.

The original position "ma[k]e[s] good" this "defect" in Kant's view by imposing the requirement. (*TJ*, p. 255) It imposes it by representing us as moral persons who are forced to choose principles from which we will act among other embodied persons in circumstances of justice which we share (see *TJ*, pp. 252-53, 257 and 515) -- a point Rawls puts more picturesquely by saying that the original position "is in important ways similar to the point of view from noumenal selves [can] *see the world*". (*TJ*, p. 255, emphasis added)

At the same time, the veil of ignorance maintains the independence of noumenal selves from the contingencies of the world they have in view. If noumenal selves are taken to inhabit the original position then, Rawls implies, Sidgwick's objection can be answered and their conditions *can* satisfy (R.1). And so the original position is choice situation which is philosophically preferred, not just to Locke's choice situation, but to Kant's as well. It is preferred to Kant's because unlike Kant's choice situation, it represents human beings as free and equal moral persons who know they must inhabit the world. Thus not only do agreements reached in the original position satisfy (R.1), but because the original position provides a "general analytic method for the comparative study of conceptions of justice" (TJ, p. 121), it gives Rawls's theory diagnostic as well as constructive power. For it enables Rawls to show why agreements reached in other choice situations fail to satisfy that condition.

§5. Satisfying (R.2)

(R.2′) requires:

> (R.2') None of the members of the well-ordered society can have sufficient reason to deviate from the principles of justice, at least so long as all the others comply with them, and all do comply by conducting themselves as "moral persons, as creatures having a conception of their good and capable of a sense of justice." (*TJ*, p. 19)

(R.2′) is a conjunction. The argument that the agreement reached in the original position satisfies it assumes what Rawls tries to show in *TJ*, chapter 8: that members of the well-ordered society all have a sense of justice, which includes a powerful and standing desire to act from principles of justice by treating them as regulative of one's conduct. The argument then proceeds in two steps, corresponding to the condition's two conjuncts.

To see how the first step of the argument goes, it is necessary to see why members of the well-ordered society want to act from principles of justice. The answer, according to Rawls, is that by acting from the principles, members of the well-ordered society express their nature as moral persons. Call this "the

expression claim". Rawls does not say exactly what he means by 'express' and its cognates. I take the notion to be adverbial and, more specifically, to characterize the way someone lives or conducts herself. To express one's nature as a moral person is to live as one. It is to exercise the features of moral personality, and thus to live in a way that is characteristic of a being who has those features. And so the expression claim says that to act from one's settled desire to comply with principles of justice is to live as such a being. Since acting from a desire to comply with the principles is the same as acting on them, we can recast the expression claim as the claim that to act on the principles is to conduct oneself as a moral person.

Rawls's defense of the expression claim so construed depends crucially on a premise which he lays down in the section on the Kantian interpretation:

> "to express one's nature as a being of a particular kind is to act on the principles that would be chosen if this nature were the decisive determining element." (*TJ*, p. 253)

The abstractness and generality of this premise make it difficult to assess, but we can get some purchase on it by thinking about the case in which Rawls is most interested, moral persons. Moral persons express their nature as such persons, or live as such persons, when they act freely. Rawls thinks that the relevant kind of freedom is autonomy, and hence that moral persons act freely when they act from principles they would give themselves. When their nature as moral persons determines the principles they choose, those principles are ones they choose or give to themselves as persons with that nature. Since members of the well-ordered society are represented as moral persons in the original position, their nature as such persons is "the decisive determining element" in the choice of principles there.

This conclusion, together with the expression claim, implies that if human beings act on the principles by regulating their conduct by them,[12] they express their nature as – and so conduct themselves as – "moral persons … capable of a sense of justice.'" This implication, in turn, gets us some way toward showing that agreements reached in the original position satisfy the second conjunct of (R.2′), which requires

[12] Another way to act on them is to express appropriate guilt for violating them. I shall leave this complication aside here.

that members of the well-ordered society comply with the principles by conducting themselves as "moral persons …having a conception of their own good and capable of sense of justice". But pursuing this implication would get us ahead of ourselves, for we first have to see whether members of the well-ordered society would act on their sense of justice and comply with the principles. To see that they would, we still need to see why the first conjunct of (R.2′) is satisfied. Though members of the well-ordered society are assumed to have a sense of justice, they might also have desires which provide them sufficient reasons to deviate from the principles. In that case, their conceptions of their good would undermine the agreement reached in the original position.

Rawls addresses this worry late in *TJ*, in section 86. His arguments are quite complicated, and I have laid them out at length elsewhere.[13] Very roughly, Rawls thinks that members of the well-ordered society all have a very strong desire to be just persons under that description. That desire belongs to their good and they can satisfy it only by regulating their plans and conduct by their sense of justice. Since the desire is strong, their conceptions of their own good would move them to comply with, rather than to deviate from, the principles. But Rawls also argues that whatever else members of the well-ordered society want, they would all also have a powerful desire to express their nature as free and equal rational beings under *that* description. That desire, too, belongs to their good. Since objects of desire are individuated by their descriptions, it has a different object than the desire to be a just person. But it follows from the expression claim that the desire to express one's nature – like the desire to be a just person -- can be satisfied only by regulating one's plans by principles that would be chosen in the original position. And since members of the well-ordered society would all have a "lucid grasp" of justice as fairness (*TJ*, p. 572), each of them would *know* that she can satisfy her desire to express her nature only by treating principles of justice as regulative, at least if others do the same. And so Rawls concludes that each member of the well-ordered society would plan to preserve and act from her sense of justice as regulative, at least if others do.

This argument is meant to show that in a well-ordered society, each person's plan to be just is what Rawls calls her "best reply" to the similar plans of others. (*TJ*, p. 568) Since a Nash equilibrium is a strategy

[13] Paul Weithman, *Why Political Liberalism?* (Oxford University Press, 2010).

combination in which each player's strategy is the best reply to the strategies employed by the other players, his use of the phrase "best reply" suggests that he thinks these plans or conceptions of the good – taken together -- are such an equilibrium. The equilibrium supports the agreement that would be reached in the original position. For when it obtains among persons with a regulative sense of justice, each affirms that the sense of justice is itself good for her. In that case, no one has sufficient reason to deviate from the demands of justice, as the first conjunct of (R.2′) requires, and so each reaffirms her plan to pursue her good justly and everyone complies with the principles, as the second conjunct requires.[14]

The second conjunct of (R.2′) requires, not just that members of the well-ordered society comply with the principles, but that they do so "by conducting themselves as 'moral persons, as creatures having a conception of their good and capable of a sense of justice.'" We have already seen that when they comply with the principles, they conduct themselves as "moral persons … capable of a sense of justice". What of the other feature of moral personality referred to in (R.2′)? Do members of the well-ordered society who treat the principles of justice as regulative comply by "conducting themselves as 'moral persons … having a conception of their good'"?

The question assumes a distinction where there is no difference. For what the principles of justice are regulative *of* is someone's pursuit of her plan of life. To treat the principles as regulative is therefore to pursue one's plan in a certain way, a way that is characteristic of a moral person. Each person's plan of life gives her conception of the good. Since members of the well-ordered society would comply with the principles agreed to in the original position by conduct which is characteristic of creatures capable of a sense of justice, they *ipso facto* comply with them by conduct characteristic of creatures with conceptions of their good as well. The agreement that would be reached in the original position therefore satisfies (R.2′).

---

[14] The last inference depends upon Rawls's answer to assurance problems. I shall leave those problems aside for reasons of simplicity and brevity, noting only the following. One of the ways members of the well-ordered society assure each other that they are committed to justice as fairness is by complying with the guidelines of what the Rawls of *Political Liberalism* called "public reason". Those guidelines would be agreed to in the original position along with the principles of justice, and the features of the original position which make it the appropriate initial situation for the adoption of the principles make it appropriate for the adoption of those guidelines as well. Thus Rawls's solution to the relevant assurance problems, like the rest of his treatment of stability, depends upon the representational function of the original position.

§6. Satisfying (R.3')

(R.3′) requires:

(R.3') The fact that the principles would be agreed to in the contract referred to in (R.1) must be what brings it about that members of the well-ordered society comply with them, as (R.2′) requires, and the connection between the hypothetical agreement and the conduct referred to by (R.2′) must itself be established in ways which treat members of the well-ordered society as "moral persons, as creatures having a conception of their good and capable of a sense of justice." (*TJ*, p. 19)

We saw in the previous section that members of the well-ordered society comply with the principles of justice, as (R.2′) requires, by acting on their desires to be just persons and to express their nature as moral persons. We will be able to see that the agreement reached in the original position satisfies the first conjunct of (R.3′) by seeing how those desires depend upon the fact that the principles would be agreed to in the original position. As I noted when I introduced (R.3′), the dependence is causal and hypothetical agreements lack causal powers. But I suggested that the *fact* that principles of justice would be the object of a hypothetical agreement could cause compliance, and that it could do so by educating citizens via institutions which satisfy the principles and which are modeled on the hypothetical agreement. I now want to elaborate on that suggestion, arguing that the institutions of a well-ordered society educate its members by causing the desires that give rise to compliance, and that they would do so in way which satisfies the second conjunct of (R.3′).

I assume the familiarity of Rawls's argument that members of a well-ordered society would develop a sense of justice, since it is the argument in Part III of *TJ* which has attracted the most critical attention. But one feature of Rawls's treatment of a sense of justice has attracted surprisingly little attention, given what I believe to be its great importance in the argument. That is is that while the sense of justice includes a desire

to comply with the principles, it also includes the aspiration to live up to various "ideals" – that is, to various conceptions of what we might be. (*TJ*, p. 473)

The ideals to which members of the well-ordered society aspire include the ideal of a just person and the ideal of someone whose conduct expresses her nature as a moral person. It was crucial to the arguments of the previous section that members of the well-ordered society desire to live up to both of these ideals and that they know they can live up to the second only by living up to the first. Yet beyond one remark which might be construed to imply that the second ideal "exercises a natural attraction upon our affections" (*TJ*, p. 478), Rawls says little that helps us understand where the desire to live up to that ideal comes from. And beyond his remark that members of the well-ordered society have a "lucid grasp" of justice as fairness, he says little that helps us understand how they know that the two ideals are coincident in their demands. To see that the agreement reached in the original position satisfies (R.3′), we need to fill in some of the details.

The most important detail is that justice as fairness is subject to what the Rawls of the *Dewey Lectures* called "the full publicity condition". (*CP*, pp. 324-26) This condition requires that all aspects of justice as fairness be available in the public culture of a well-ordered society. The original position, the ways in which it represents the features of moral personality, and the necessity of acting from principles chosen there for the expression of our nature as moral persons, would all be "publicly available" to members of the well-ordered society. (*CP*, p. 324)

I believe Rawls thinks that as result of their public availability, the basic ideas of justice as fairness will permeate casual political discussion in a well-ordered society, as talk of natural rights pervades casual discussion in our own. But I assume he also thinks that these ideas will be appealed to systematically by public officials to justify policies and court decisions which are just and are publicly perceived to be so. These appeals will familiarize members of the well-ordered society with the ideal of themselves as free and equal moral persons who can live as such only by complying with political outcomes that would be arrived at in an appropriate situation – that is, in a situation in which the defining features of our nature are appropriately represented. The justice of those decisions makes the underlying ideal attractive, and elicits a desire to live up to it.

The claim that this is what Rawls thinks – and the dependence of this line of thought on the representational function of the original position -- can be substantiated by Rawls's discussion of the points of view from which political decisions are to be made and court cases to be decided. In section 31 of *TJ*, Rawls introduces the "four-stage sequence" for arriving at and implementing his two principles. The first stage is the original position, the second is a constitutional convention, the third is the "legislative stage" and the fourth is the judicial stage. The stages form an ordered sequence because the conditions of later stages are specified with reference to those which precede it. Progressively more information is made available to parties at each succeeding stage, and parties at later stages are faced with choice-problems which are constrained by the choices made at earlier ones.

The original position clearly enjoys a conceptual priority in the definition of the sequence. Since later stages can be viewed as its "descendants", the original position gives the sequence a form of unity we might call "genetic". But it is not clear that Rawls gives the original position priority in virtue of its representing moral personality. Even if he does, it might seem clear that later stages in the sequence cannot "inherit" this representative function, since parties at later stages are not represented *just* as free and equal moral persons since they have access to contingent information about themselves and their society. Thus the four-stage sequence may, as Rawls says, be "a device for applying the principles of justice." (*TJ*, p. 200), but it seems not to be a device *of representation*. And while I have contended that the original position gives Rawls's justice theoretical unity in virtue of its representative function, that seems not to be true of its role in defining the sequence.

But an important remark late in *TJ* suggests the opposite conclusion. At the end of the difficult section on the "Unity of the Self", Rawls says that "the notion guiding the entire construction [of the four-stage sequence] is the original position and its Kantian interpretation." (*TJ*, p. 566) The remark is part of an argument for Rawls's claim that principles of justice and the rules that implement them do not establish a dominant end, but rather "approximate the boundaries" within which people are free to choose plans of life. (*TJ*, p. 566) The implication of this claim is clearly supposed to be that while principles of justice, the constitution, laws and judicial decisions of a well ordered society may all limit human conduct, in justice as fairness they do so in ways that are consistent with human freedom.

If freedom is understood as "autonomy", as I believe it must be, then this can only be true if at each stage, the rules adopted are in some sense ones we give or would give ourselves. They might not be ones we would give ourselves as free and equal rational beings – for while our nature as free and equal rational beings might serve as the "decisive determining element" in the original position (*TJ*, p. 253), it is too indeterminate to guide the choice of legislation. But they are ones we would give ourselves as free and equal authors of a constitution or free and equal legislators. If this is correct, then the core ideas of freedom and equality are adapted to the demands of each stage in the four-stage sequence.[15] The adaptation is *of* those ideas *as modeled in the original position*. That is why Rawls says that the original position "guid[es]" the construction of the sequence. So the original position is given definitional priority in that sequence, and gives genetic unity to it, because it is a device for representing moral personality. The use of that device by public officials, public knowledge of its use, and its availability to citizens who want to think through political and judicial questions for themselves (*TJ*, p. 473), all educate citizens in the ideals of justice and moral personality that it represents. That education is a moral one, broadly understood: members of the well-ordered society incorporate that ideal into their conceptions of the good as an end which is desirable for its own sake. [16]

Rawls sums up this process of moral education near the end of his second *Dewey Lecture*, saying that when the full publicity condition is satisfied:

> The derivation of citizens' rights, liberties and opportunities invokes a certain conception of their person. In this way citizens are made aware of and educated to this conception. They are presented with a way of regarding themselves that otherwise they would most likely never have been able to entertain. Thus the realization of the full publicity condition provides the social milieu within which the notion of full autonomy can be understood and within which its ideal of the person can elicit an effective desire to be that kind of person. (*CP*, pp. 339-40)

[15] As if to confirm this, Rawls says that "the constitutional process should preserve the equal representation of the original position to the degree that this is practicable." (*TJ*, p. 222)

[16] On the psychological laws as governing the acquisition of new final ends, see *TJ*, pp. 493-94.

The argument which I have imputed to Rawls in this section, and which is summed up in the passage I have just quoted, shows that the agreement which would be reached in the original position satisfies the first conjunct of (R.3′). Moreover, in *TJ*, §78 Rawls offers what is, in effect, a sequence of very powerful arguments for the conclusion that the second conjunct is satisfied as well. He says that in the well-ordered society, each person's "moral education itself has been regulated by the principles of right and justice he would consent to in an initial situation in which all have equal representation as moral persons." (*TJ*, pp. 514-15) He offers further argument for this conclusion in the *Dewey Lectures* when he defends the condition of full publicity. There he argues that the condition is "fitting" because a society's conception of justice shapes the character of its citizens, and this can be consistent with treating them as free and equal moral persons only if they can examine the conception and understand its effects on them. (*CP*, pp. 325-26) If these arguments are sound, as I believe they are, then the agreement reached in the original position satisfies (R.3′) as well as (R.1) and (R.2′). That agreement is therefore *Rawlsian self-enforcing*.

§7. Taking Stock

Let us take stock. Showing that the agreement that would be reached in the original position satisfies (R.1) required showing that the original position is the appropriate device for identifying principles of right. Showing that that agreement satisfies (R.2′) required showing that members of the well-ordered society could satisfy desires which are central to their good only by acting from principles that would be adopted in the original position. Showing that it satisfies (R.3′) required showing that facts about what would be agreed to in the original position themselves help to elicit those desires and to do so in an acceptable way.

All these "showings" depend upon the representative function of the original position. For what makes the original position the appropriate device for identifying principles of right is the fact that it represents the features of moral personality. What makes acting from the principles of right good for members of the well-ordered society is that they can satisfy the desire to express their nature only by acting on principles which would be chosen in the original position. As we saw, the argument for *that* conclusion

depends upon the claim that the features of moral personality are the "decisive determining element" of the agreement in the original position. They are the decisive determining element of that agreement because the original position represents us simply as moral persons. Finally, what elicits the desire to be just and to express our nature is the public use of the original position -- as a device of representation – to arrive at and justify just political outcomes.

Thus once we see the original position as a device of representation, and understand the many arguments in which Rawls deployed it, we can see how precisely he contrived that device to "bridge" the right and the good. He bridged the right and the good in order to connect justice with stability, thereby using a single versatile device to join seemingly different subject matters within the embrace of a single grand theory. Dispensing with the original position would deprive that theory of its methodological unity. It would also deprive the theory of considerable power. Not only would justice as fairness would lose the diagnostic power to which I referred at the end of §4, but it would also lose a device which can clearly represent our nature in the public political culture. It would therefore lose the educational power which maintains the justice of the well-ordered society over time.

I have argued that the well-ordered society would be stably justice because the agreement that would be reached in the original position is a Rawlsian self-enforcing agreement. Rawls's connection of justice and stability is philosophically important because Rawlsian self-enforcement is an interesting and desirable property for an agreement to have, even if the equilibrium concept on which it depends is not quite as demanding as we might like.[17] For it might seem, and has seemed to many thinkers in the history of political philosophy, that the benefits of social cooperation can be had only at the cost of freedom. Hobbes seemed most pointedly to have raised this worry by arguing that the agreement to enter into a social order can be sustained only if the parties to the agreement also agree to submit themselves to an absolute sovereign.

Self-enforcing agreements are attractive because they are honored without a mechanism of enforcement. Instead, parties to these agreements are moved to honor them by the payoffs of doing so, which they regard as superior to the payoffs of deviation. And so as I noted earlier, agreements which satisfy the

[17] See the considerations adduced in note 7 above.

conditions on self-enforcement are honored freely, in some sense of 'free'. How robust or valuable that kind of freedom is depends upon how the payoffs move parties to the agreement and how the parties have come to value those payoffs.

The conditions of Rawlsian self-enforcement strengthen the conditions of self-enforcing agreements by adding significant moral demands. As a result, agreements which are Rawlsian self-enforcing are upheld by parties who value being just persons and expressing their nature, and who regard payoffs they could gain from violating the principles as being "without value." (*CP*, p. 106). Members of the well-ordered society have the payoff structure they do because they have been educated to be persons with those values, and educated by processes which are, and which they can see to be, consistent with their freedom and equality. And so agreements which satisfy the three Rawlsian conditions are upheld by parties who conduct themselves as free and equal moral persons. Because of the conditions of the original position, the kind of freedom they express in their conduct is freedom of an especially robust kind. For in regulating their pursuits by principles that they would adopt there, members of the well-ordered society express their autonomy. The elegant connection Rawls forged between justice and stability therefore makes it possible to answer Hobbes and to characterize the enduring form of political association that Rousseau sought – one "in which each while uniting himself with all, may still obey himself alone, and remain as free as before."[18]

§8. The Turn to Political Liberalism

Rawls famously opens section 1 of *TJ* by observing that "justice is the first virtue of social institutions". He continues immediately "as truth is of systems of thought. A theory however elegant and economical must be rejected or revised if it is untrue[.]" The remark was prescient. For while the theory of justice Rawls laid out in *TJ* was elegant and economical, he came to believe that not all its claims could be true, since it suffered from an inconsistency between "the account of stability given in part III of *Theory*" and "the view as a whole." (*PL*, pp. xvii-xviii)

[18] Jean-Jacques Rousseau, *The Social Contract*, chapter vi.

In brief, the inconsistency is this. As saw in the section on (R.2′), Rawls's argument for stability depends on his claim that everyone in the well-ordered society would want to express her nature as a moral person by conducting herself autonomously. And as we saw in the section on (R.3′), Rawls thought that that desire would be elicited by public knowledge and institutionalization of the agreement reached in the original position. But that agreement supports free institutions, and free institutions encourage those who live under them to adopt different views of the kind of persons they want to be. And so while the "account of stability given in part III of *Theory*" requires that members of the well-ordered society converge on the same partial conception of the good, "the view as a whole" encourages them to diverge.

Since Rawls was not going to give up his defense of free institutions or his claim that they encourage moral pluralism, the inconsistency in justice fairness required him to give up his claims about moral convergence. These are claims Rawls relied upon to argue that the agreement which would be reached in the original position satisfies (R.2′) and (R.3′), and so would be Rawlsian self-enforcing. The inconsistency therefore forced Rawls to think anew about whether members of the well-ordered society could uphold that agreement while conducting themselves autonomously.

Rawls's response to the inconsistency was, of course, to re-present justice as fairness as a political liberalism. This change began with changes in the basic intuitive ideas on which justice as fairness was said to rely and in what the original position was said to represent. Whereas the Rawls of *TJ* has relied on a basic intuitive idea of moral personality and used the original position to represent its contents, the later Rawls founded justice as fairness on a basic intuitive idea of political personality or citizenship. Since the features he said are definitive of citizenship are those he had previously used to define moral personality – namely, the capacities for a conception of the good and for a sense of justice (*CP*, p. 398) -- he could claim that the original position represented the contents of *this* basic idea instead. (See *PL*, pp. 103-4)

We have seen how the Rawls of *TJ* used the original position to connect justice and stability. The later Rawls kept that connection in place, but the change in what he used the original position to represent allowed him to connect them differently and less controversially than he had in *TJ*. Instead of arguing that principles of justice are those which would be agreed to by free and equal persons, he identified them with those that would be agreed to by free and equal citizens. (*CP*, p. 399) And instead of arguing that the

agreement reached in the original position would enforce itself because everyone in the well-ordered society would develop a sufficiently strong desire to express her nature as a moral person, he argued – in effect -- that it would enforce itself because everyone would develop a sufficiently strong desire to live up to the conception of citizenship the original position was now said to represent. But to see how the new arguments for self-enforcement go, and to see that they are less controversial than the arguments of *TJ*, we need to see how the conditions of self-enforcement – and the arguments for their satisfaction – change as a result of the turn to political liberalism.

Since the later Rawls identified principles of justice with principles that would be accepted by those who live under them, now conceived as free and equal citizens, he can be read as endorsing what we might call a "political variant" of (R.1):

> (PL.1) The principles of justice that members of the well-ordered society are to follow must be specified by the terms of an agreement or contract among them, conceived as citizens having a conception of their good and capable of a sense of justice.

Because the powers definitive of citizenship are those which were previously said to be those definitive of moral personality, and because the original position represented the latter, it also represents the former. Many of the arguments canvassed above in section 4, which show that the agreement reached in the original position satisfies (R.1), can be re-purposed to show that it satisfies (PL.1). The exceptions are the arguments which purport to show that the superiority of the original position to Kant's interpretation of situation in which all moral principles are arrived at – namely, the situation of noumenal selves. For that argument seems to depend upon the use of the original position to represent us having the features of moral personality in virtue of which those principles are to bind us, and not just the features of citizenship in a liberal democracy.

(R.2′) and (R.3′) are the conditions on self-enforcement which the agreement reached in the original position cannot satisfy because of the inconsistency Rawls found in *TJ*. And so as we would expect, more

interesting and significant departures from the arguments for Rawlsian self-enforcement come to light when we ask whether the agreement reached in the original position would satisfy their political variants:

> (PL.2') None of the members of the well-ordered society can have sufficient reason to deviate from the principles of justice, at least so long as all the others comply with them, and all do comply by conducting themselves as citizens having a conception of their good and capable of a sense of justice.

and

> (PL.3') The fact that the principles would be agreed to in the contract referred to in (PL.1) must be what brings it about that members of the well-ordered society comply with them, as (PL.2′) requires, and the connection between the hypothetical agreement and the conduct referred to by (PL.2′) must itself be established in ways which treat members of the well-ordered society as citizens having a conception of their good and capable of a sense of justice.

Let us start with (PL.2′). The later Rawls, like the earlier, would maintain that members of the well-ordered society would all develop a sense of justice by the processes laid out in *TJ*, chapter 8. Indeed, I believe he never revisited the arguments of that chapter after his turn to political liberalism because he thought the arguments were successful as they stood. As we saw in §5, the sense of justice is – in effect though not in intention -- a desire to act from principles chosen in the original position. As I have stressed, the later Rawls used the original position to represent the defining features of citizenship. If he continued to endorse what I referred to in §5 as “the expression claim”, as I believe he did, then he could conclude that by

acting from the principles, members of the well-ordered society can express their nature as citizens capable of a sense of justice.[19]

As we also saw in §5, Rawls moved from that conclusion to the further conclusion that the agreement reached in the original position satisfied both conjuncts of (R.2′) by showing that when all the members of the well-ordered society plan to regulate their pursuits of their good by the principles of justice, their plans are in a Nash-type equilibrium. The argument for *that* conclusion depended upon each person's attaching a very high value to the expression of her nature as a moral person. The later Rawls could move, by parity of reasoning, to the conclusion that the agreement reached in the original position satisfies (PL.2′) if he could argue that members of the well-ordered society would all value the expression of their citizenship highly enough that it would outweigh other elements of their good. Since the conception of citizenship represented in the original position is strictly political and does not apply to the whole of life, the new argument would not depend upon convergence on a comprehensive conception of the good – even a partially comprehensive one. But it does require convergence on a political conception of justice. Rawls does seem to think that such convergence would come about, and that all members of the well-ordered society "would want to realize in their person, and have it recognized that they realize, that ideal of citizens." (*PL*, p. 84) But under what conditions would that convergence come about? And under what conditions would they attach enough value to expressing their nature as citizens to reach the conclusion Rawls needs?

Rawls's answer is that they would do so if an overlapping consensus obtains. For when an overlapping consensus obtains, "[r]easonable conceptions [of the good] endorse the political conception, each from its own point of view." (*PL*. p. 134) That is, when an overlapping consensus on justice as fairness obtains, reasonable comprehensive doctrines all recognize its values, including the value of living up to its ideal of citizenship. If members of the well-ordered society all follow their comprehensive doctrines, as we assume they do, then they all "will judge (by their comprehensive doctrine) that political values" – including the value of living as a free and equal citizen – "outweigh or are normally (though not always) ordered prior

[19] It might be thought that artifacts -- and therefore social roles, such as citizenship – cannot have natures, and that we should not read Rawls supposing that they do. Note that at *PL*, p. 203, Rawls says that "having the two moral powers" belongs to "the essential nature of citizens."

to whatever non-political values may conflict with them." (*PL*, p. 392) In that case, each person's plan to live up to the ideal of citizenship – as represented in the original position – is her best reply to the similar plans of others. A Nash-type equilibrium obtains, each regulates her pursuit of her conception of her good by her sense of justice, and the agreement that would be reached in the original position satisfies (PL.2′).[20]

What of (PL. 3′)?

When I reconstructed Rawls's argument that the agreement reached in the original position would satisfy (R.3′), and would do so in virtue of the representational function of the original position, I devoted most of my attention to an argument that it would satisfy the first conjunct. Now it is the second conjunct that merits most of the attention. For the argument that the first conjunct would be satisfied depended upon showing that members of the well-ordered society would all come to aspire to two ideals: the ideal of a just person and the ideal of someone whose life expresses her moral personality. The turn to political liberalism meant a shift from talk of the latter ideal to talk of an ideal of citizenship.

But despite this important shift, the conclusion that the agreement reached in the original position would satisfy the first conjunct in (PL.3′) would be much the same as the argument that it would satisfy the first conjunct of (R.3′). For the later argument like the earlier one depends upon claims about the education effected by the full publicity condition and about the educational value of adopting the different points of view in the four-stage sequence.[21] Those are claims I believe Rawls continued to endorse even after his turn to political liberalism. As if to confirm that he did, Rawls repeats the summary quote from the original *Dewey Lectures* that I cited at the end of section 6 is repeated virtually word for word in the revised version of the *Dewey*'s which appears in *Political Liberalism*. The only difference which the shift made is that Rawls replaced the sentence

[20] See note 14.

[21] For the latter, see *PL*, pp. 397-98.

> Thus the realization of the full publicity condition provides the social milieu within which the *notion of full autonomy* can be understood and within which its ideal of the person can elicit an effective desire to be that kind of person. (*CP*, p. 340, emphasis added)

with:

> To realize the full publicity condition is to realize a social world within which *the ideal of citizenship* can be learned and may elicit an effective desire to be that kind of person. (*PL*, p. 71, emphasis added)

Of course, members of the well-ordered society will comply with the principles of justice only if the desire to live up to the ideal of citizenship is, as Rawls says here, "effective". Showing that the agreement reached in the original position would satisfy the second conjunct of (PL.3′) requires showing that the process by which the fact of agreement in the original position educates members of the well-ordered society into a desire which is effective must itself treat them as citizens. We saw above that that desire will be effective if – and perhaps only if -- an overlapping consensus obtains and reasonable comprehensive doctrines support that ideal. So to show that that the agreement reached in the original position satisfies the second conjunct of (PL.3′), Rawls would need to show that the fact that the principles would be agreed to helps to bring about an overlapping consensus in a way that satisfies the requirement that conjunct expresses. How, then, does an overlapping consensus come about? Does its coming about depend upon the representative function of the original position?

Rawls conjectures that many members of the well-ordered society will endorse comprehensive doctrines which "are not seen by them as fully general and comprehensive". (*PL*, p. 160) These citizens will not see that their religious or philosophical views about the most worthwhile life have many clear implications for politics, or at least any clear implications for which conception of justice they should endorse. Rawls implies that an overlapping consensus obtains when reasonable comprehensive doctrines "understand the wider realm of values to be congruent with, or supportive of, *or else not in conflict with*,

political values as these are specified by a political conception of justice[.]" (*PL*, p. 169, emphasis added) The "slippage" between comprehensive doctrine and political conception (*PL*, p. 160) opens the possibility that many members of the well-ordered society can view their comprehensive doctrines as participating in an overlapping consensus simply by virtue of their doctrines' consistency with justice as fairness. Rawls conjectures that these citizens will acquire desires to be just and to express their citizenship from the public culture, and that those desires will be effective simply because their conceptions of the good are not comprehensive enough to encourage political ideals which might conflict with them. (See *PL*, p. 168) Since the educational processes of the public culture treat members of the well-ordered society as citizens with the two moral powers, the agreement reached in the original position satisfies (PL.3′) in these cases.

More complicated questions are raised by citizens who *do* see their comprehensive doctrines as "fully general and comprehensive". These comprehensive doctrines include ideals of political conduct and they encourage their adherents to desire to live up to those ideals. These comprehensive doctrines can take part in, and be seen to take part in, an overlapping consensus on justice as fairness only if they are "congruent with, or supportive of", and are known to be "congruent with or supportive of", its political values. The question of how an overlapping consensus comes about, and whether it comes about in a way that satisfies the second conjunct of (PL.3′), are questions about how fully and generally comprehensive doctrines come to be "congruent with or supportive of" justice as fairness.

At one point, Rawls ventures that this can happen because "A reasonable and effective political conception may bend comprehensive doctrines toward itself, shaping them if need be from unreasonable to reasonable." (*PL*, p. 246) But he says little about the processes by which this bending might take place. No doubt he says little because the processes depend upon the complex cultural and sociological forces which would be at work in a just society. These processes are not well understood, they are likely to operate on different comprehensive doctrines in different ways, and understanding them lies outside the subject-matter and expertise of political philosophy.[22]

[22] I discuss them in *Why Political Liberalism*, pp. 308-12.

What matters for present purposes is that doctrines which are comprehensive enough to require such shaping are articulated in comprehensive systems of thought. Their content and implications develop – and so are "ben[t]" -- through the intellectual work of their theoretically-inclined adherents. Those adherents, like all other members of the well-ordered society, are attracted to justice as fairness by the workings of a public culture which satisfies the full publicity condition and which treats them as free and equal citizens. Their education, the formation of their views, and the arguments they offer, all go some way toward sustaining the causal connection from the fact that the principles of justice would be agreed to in the original position – via an overlapping consensus -- to compliance with the principles by adherents of their comprehensive doctrine. It therefore goes some way toward insuring that the agreement reached in the original position satisfies (PL.3′).

But compliance by adherents of fully general comprehensive doctrines is not just brought about the arguments of their intellectual leaders. It is also brought about by their education in their doctrines – including, perhaps, education by authority. That education may not enable them to realize the full autonomy of moral persons as initially modeled in the original position. It must, however, treat them as free and equal citizens. (*PL*, pp. 199-200) And so the way in which the agreement reached in the original position enforces itself in political liberalism has to be qualified accordingly. That qualification is reflected in the kind of freedom that members of the well-ordered society realize when they comply with the principles of justice.

§9. Conclusion

Rawls's turn to political liberalism leaves much of his original treatment of justice in place. A merit of the reading presented here is that it documents the continuity between Rawls's earlier and later work by showing how many of the arguments for (R.1), (R.2′) and (R.3′) can be used, mutatis mutandis, to support (PL.1), (PL.2′) and (PL.3′). As we have seen, those arguments depend – indeed I think they ineliminably depend -- upon the representative function of the original position and upon the educational effects of publicizing and institutionalizing the agreement reached in it. Because those arguments use the original position to bridge the right and the good, and thereby to connect justice and stability, Rawls's revised treatment of justice retains the scope and methodological unity of its initial presentation in *TJ*. It does so

despite the fact that Rawls changed what he devised the original position to represent. Even recast as a political liberalism, Rawls's treatment of justice remains grandly theoretical.

When I introduced (R.1), (R.2′) and (R.3′), I said that they are jointly sufficient for Rawlsian self-enforcement. I did not say that they are necessary, for I take (PL.1), (PL.2′) and (PL.3′) also to be jointly sufficient: because the agreement that would be reached in the original position satisfies them, that agreement enforces itself. I have said that the attraction of a self-enforcing agreement is that it is upheld by the free conduct of those who are party to it, and I noted that when an agreement satisfies (R.1), (R.2′) and (R.3′), those who uphold it realize freedom of an especially robust kind. In the original *Dewey*'s, Rawls called that kind of freedom "full autonomy". (*CP*, p. 340) An agreement which satisfies (PL.1), (PL.2′) and (PL.3′) is not upheld by conduct which realizes such autonomy, since the conduct by which members of the well-ordered society uphold it does not express their nature as moral persons. Rather, the principles which they would agree to in the original position are principles they would give themselves as political persons, and they comply with the principles by conducting themselves as such. In complying with an agreement on those principles which is self-enforcing, they realize an autonomy which Rawls was still willing to describe as "full" but which he insisted is "political not ethical". (*PL*, p. 77)

Paul Weithman

Department of Philosophy

University of Notre Dame