THE PROBLEM OF KNOWLEDGE IN THE ORIGINAL POSITION*

Margarita Levin

Much has been written about John Rawls' A Theory of Justice.[1] One of its most distinctive features is the "original position" in which principles of justice are chosen. In this paper I discuss the formulation of the original position and, in particular, the conditions of knowledge that Rawls puts on it. I will also discuss the relationship between the original position and what I shall call the "entry condition." I will argue that it is very difficult, if not impossible, to satisfy all of Rawls' requirements for the original position. I end with a discussion of changes that could be made in the system and the effects they would have.

I

Before getting down to details, I will sketch briefly that part of Rawls' theory which bears on the issues discussed in this paper. Rawls labels his theory "justice as fairness." By this he means that his principles of justice are those that would be chosen by "free and rational persons concerned to further their own interests . . . in an initial position of equality as defining the fundamental terms of their association" (p. 11). The initial position of equality that Rawls specifies--the "original position" --is Rawls' device for capturing our intuitions about the conditions under which principles of justice should be chosen. Rawls stresses that the original position should be "interpreted so that one can at any time adopt its perspective" (p. 139). Because the particular requirement comes up so often, I will refer to it as the "entry condition." We shall see below that adherence to the entry condition poses considerable problems for Rawls.

There are, furthermore, certain formal constraints on the conception of justice that Rawls finds intuitively "natural enough" (p. 131).[2] The first is that principles be general, that they not contain

---

*I would like to thank Allen Buchanan, Norman Dahl, and my husband Michael for many helpful suggestions.

proper names or definite descriptions. Second, "[p]rinciples must be universal in application" (p. 132). Principles must also be public; they must be known by all, not just implicitly followed. They must be competent to adjudicate conflicting claims. Finally, there can be no appeal beyond these principles.

The chief characteristic of the original position, however, is the limited knowledge of the parties involved. They are supposed to know general facts about economics, psychology and sociology, but no specific facts about themselves. They do not know their own status, their talents and liabilities, their sex, race, physique or anything else. This "veil of ignorance" is intended to ensure that the people in the original position--henceforth "the participants"--do not choose principles that are specially tailored to their particular circumstances. For instance, if a participant knew he were white, he might opt for principles that would permit the enslavement of Blacks. The absence of personal information ensures that every level of society gets a fair deal, since the participants realize that when the veil is lifted they may find that they belong to the lowest level. (One might see a certain overlap between the first constraint and the information gap: both rule out principles partial to individuals.)

Rawls claims that the particular circumstances of the original position will lead the participants to choose certain definite principles of justice. The first, the greatest equal liberty principle, is that everyone is to have as much liberty as is compatible with everyone else's having the same amount of liberty. The second principle, the one the veil of ignorance is designed to elicit, comes in two parts. First of all, there is to be fair equality of opportunity--everyone is to have a fair chance at positions of power and prestige. This is intended to neutralize the disadvantages that one's family or social position (or abilities bestowed by the "natural lottery") may create. The second clause is known as the difference principle: "social and economic inequalities are to be to the greatest benefit of the least advantaged" (p. 302). If the participants adopt this principle, then even if they turn out to be among the least advantaged in their society they will take comfort in the thought that any other arrangement of inequalities would have left them worse off. Rawls argues that the choice of the difference principle will be the result of the participants' using the maximin strategy for choice under conditions of uncertainty: choose from among a set of alternatives

the one whose worst outcome is the best it can be. Rawls argues at great length that the maximin strategy is the appropriate one for the participants to use behind the veil of ignorance.

I will not discuss whether maximin is indicated under the circumstances, nor whether its use will indeed lead to Rawls' two principles. The coherence of the original position is logically prior to these issues; if it proves unworkable these other problems lose their interest.

II

The original position plays a crucial role in Rawls' theory. Its purpose is to "make vivid to ourselves the restrictions it seems reasonable to impose on arguments for principles of justice, and therefore on these principles" (p. 18). Rawls intends to show that our intuition about the rules that should guide the choice of principles do agree with the set-up of the original position, and that his principles of justice are the ones that the participants whould choose. Given that result, he then argues that since these are the principles we would choose if we were in the original position, these are the principles we should in fact adopt. This is an argument from "hypothetical consent." Rawls intends his situation to be analogous to the Kantian one of a purely rational being autonomously legislating for himself. Just as Kant argues that we should follow those rules a rational being would choose, so Rawls argues that we should adopt those principles the participants would choose. While Rawls never, to my knowledge, explicitly characterizes his argument in this way, some such tacit assumption must be at work. If not, the moral significance of what the participants would do becomes obscure.[3] Clearly, then, if Rawls' argument is to persuade us, the original position must perform as advertised. It must be possible for us to imagine the situation with the entry condition simultaneously fulfilled--the knowledge conditions must not prevent our adopting the perspective of the original position.

There are two possible interpretations of the "entry condition." The weaker one is that it simply require the original position to be a coherent concept, that it be possible for us to imagine the original position set up and to see that the principles of justice do indeed follow from it. On this reading, the requirement that we be able to "at any time adopt its perspective" just means that we should be able to understand what is going on there. But this is surely

a rather unlikely reading. It goes without saying that the original position should be coherent and understandable. This is a requirement every theory must meet. Rawls would not pay such special attention to this requirement were that all he meant. Therefore, the stronger reading must be the correct one, namely that we must be able to imagine ourselves entering the original position now. In practical terms, that amounts to our arguing for principles of justice just on the basis of the information that would be available to us if we were in the original position.

What I have called the entry condition is not to be confused with what Rawls calls the "present time of entry interpretation." This, judging from where it appears (in sections dealing with saving for future generations), is the stipulation that the participants are all contemporaries, and that they know this. That is, the participants are not to be thought of as coming from different centuries: "Since the persons in the original position know that they are contemporaries (taking the present time of entry interpretation) . . ." (p. 140). Because they know they all belong to the same time, the participants need make no provision for saving. It must be stipulated that future generations are to be considered to prevent the participants from choosing principles that are intuitively unjust.

III

Since the original position is nothing but the restrictions and requirements on knowledge it imposes, it is these conditions that must be examined. First let us recall the information hidden from the participants by the veil of ignorance:

> [N]o one knows his place in society, his class position or social status; nor does he know his fortune in the distribution of natural assets and abilities, his intelligence and strength, and the like. Nor, again, does anyone know his conception of the good, the particulars of his rational plan of life, or even the special features of his psychology such as his aversion to risk or liability to optimism or pessimism. More than this, . . . the parties do not know the particular circumstances of their own society. (p. 137)

The veil does not shut out all information, however. The participants ". . . know the general facts about human society. They understand political affairs and

the principles of economic theory; they know the basis of social organization and the laws of human psychology" (p. 137). The only personal information they have is that they have "some rational plan of life" (p. 142) but they have no idea what their particular ends or interests may be.

To begin with, these restrictions certainly strike one as peculiar. As Rawls remarks: "Surely, some may object, principles should be chosen in the light of all knowledge available" (p. 139). It seems irrational to deny information to the decision makers. But Rawls would say that all the relevant information is available. The participants are only denied that information that might tempt them to choose principles that favor them. To some extent, this denial may be unnecessary because, as noted, the formal constraints on the concept of justice already rule out many types of tailor-made principles. For instance, there seems to be no reason why the participants should not know their own names. The generality requirement by itself ensures that a participant could not argue for a principle that uses his own name.

A further, and critical, defect is that Rawls is unclear about whether the participants lack those properties hidden by the veil or whether the participants are merely ignorant of possessing them. The former interpretation is unlikely. Under it, the participants are only stick figures with little resemblance to us and of less interest. The argument from hypothetical consent draws its strength from the claim that we can imagine ourselves in the original position and agree that we would reach the same conclusion as the participants. Compare discussions of the motives and actions of fictional characters. When we ask what Hamlet would have done in a certain situation, we are thinking of him as if he were a human being with all the traits human beings have. We could not speculate about him if we assumed him to be only the sum total of the words and actions assigned to him by Shakespeare. Similarly, if the participants' decisions are to influence us in any way we must think of them as more than just disembodied voices that decide on principles. We must think of them as people with bodies, talents, fears, ambitions, etc.: the latter interpretation is the one Rawls needs. The need to satisfy the entry condition is a parallel reason for the participants actually to possess the properties in question. Since we must be able to adopt the perspective of the original position, and we certainly have these properties, the most that can be required of the participants is that they forget they have these

properties.

But this admission undermines Rawls' derivation of his principles, because ignorance of a trait does not prevent its operation. Most people do not know how shy or optimistic or agile they are, yet these traits shape their beliefs and behavior. The gambler at the race track does not know his "aversion to risk," yet his attitude influences the bet he makes. (Someone might object that "reason and emotion are distinct." This is undeniable; but it is also obvious that these two faculties influence each other. This latter point is all the present argument requires.) The example of the gambler is particularly relevant to Rawls theory, since the participants are supposed to use the maximin strategy which in turn leads them to choose the Rawlsian principles of justice. As I said earlier, I will not go into the details of Rawls' argument for maximin. I merely want to point out that Rawls says his principles "are those a person would choose for the design of a society in which his enemy is to assign him his place . . ." (p. 152). True, he goes on to say that "the [participants] do not, of course, assume that their initial place in society is decided by a malevolent opponent. . . . [T]hat the two principles of justice would be chosen if the parties were forced to protect themselves against such a contingency explains the sense in which this conception is the maximin solution" (p. 153). In effect, Rawls claims that the participants will not, in the absence of all pertinent information, assign the same probability to all possible outcomes and act accordingly; rather, they will use the strategy prescribed for those with infinite aversion to risk. Yet he continues to emphasize that because of the veil of ignorance "the parties do not know whether or not they have a characteristic aversion to taking chances" (p. 172). My point is that whether they know it or not is irrelevant--they have the attitudes toward taking chances they do, and (if they are enough like us to represent us) no constraints on their knowledge will prevent their attitudes from influencing them. That Rawls thinks that forbidding the participants to have knowledge of their own traits is enough to secure impartiality indicates the very confusion I have cited--that between knowledge of one's personality and the effect of one's personality on one's decisions. To remedy the situation Rawls would have to claim or argue that our traits do not influence our thinking--a claim which is, to say the least, counterintuitive.

It might be objected that the participants' characteristic attitudes are not supposed to enter into the arguments for the principles of justice. The

entry condition amounts to saying that we cannot appeal in our arguments to specific information about individuals; our arguments are supposed to keep to general facts about society. This might remedy the situation if the derivation of the principles of justice were fully rigorous. However, Rawls appeals to intuition so frequently in his account of what the participants would do that he opens the door for many non-deductive types of reasoning. This is especially true in the case of something as deep-seated as one's attitude toward risk. What one person will see as a quite reasonable chance to take, given the high pay-off, another will view as foolish gambling. All the veil of ignorance rules out are arguments beginning "Because I am so timid, I think . . . ."

A broader issue raised by the distinction between having a trait and knowing that one has it is the question of how much we are influenced by our sex, race, age, strength, health and the like. These are all items the veil of ignorance hides from the participants and, because of the entry condition, from us --yet such items influence our attitudes and behavior for better or worse. To most, the fact of this influence will seem obvious. Those who disagree must at least grant that Rawls would have to argue that this is not so. The very absence of any discussion of this issue is indeed further evidence that Rawls confuses not knowing about a trait with not having it. The traits in question are pertinent because Rawls says that the participants must not be unrealistic idealists who choose principles that are impossible to abide by: "they will not enter into agreements they know they cannot keep, or can do so only with great difficulty . . . along with other considerations they count the strains of commitment" (p. 145). That is, they must bear in mind how heavy will be the burden, and how much such personality traits as envy, sloth and greed may affect one's commitment to a particular system of rules. It is not enough to assure us that once the correct principles are in effect these character traits will disappear. That sounds too much like the Marxist who airily asserts that, come the revolution, greed and ambition will disappear. In light of this, Rawls might say that the participants know enough general facts about psychology to be able to predict which principles will be difficult to abide by and which will not. But psychological theories must take into account the differences among people. Given any social scheme there will always be some people who find it difficult to fit in, to abide by its rules. So, at least, it has always been. The participants may not know whether they are among the misfits or not, but they may instinctively feel an aversion to certain

principles.

Before turning to the problems raised by the requirements on knowledge, I want to make one last point about the restrictions. The restrictions are in force to ensure that every type of person in a society will be considered when the principles of justice are being chosen. Since a participant is unaware of his race, sex, age and wealth, he must bear in mind what effect the principles will have on him no matter what he turns out to be. Rawls hopes to cover every sort of person this way, but he fails to cover them all equally. As Hare remarks, "it is rather hard to see how the veil of ignorance could conceal from one of the contracting parties the fact that he is a babe in arms."[4] Rawls writes as if it were possible for the participants to think that they might be children (see pp. 248-49). But this is impossible. The participants know psychology and they are rational, so they can draw the neo-Cartesian inference: I reason, therefore I am adult (or, at least, not a young child). If Rawls will not permit them even such inferences as this, the participants' status as our representatives is again threatened. In any event, Rawls is inconsistent: his efforts to block the formation of coalitions (see, e.g. p. 140) indicates that he does expect the participants to use their wits.

Just because children cannot be participants does not mean their rights will go unconsidered. The participants should certainly take into account the fact that they might turn out to be parents and care very much what happens to their children. But this concern cannot carry as much force as the concern for oneself and the situations one might turn out to be in. Rawls, we saw, invokes the requirement that the participants care about their descendants' welfare and what those descendants think of their decision (see p. 155). But even if this seemingly arbitrary requirement is imposed, it is clear that they cannot care for their descendants as much as they care for themselves, especially since they realize that they may not have any descendants. So Rawls' knowledge conditions do not ensure the rights of children to the same degree as they are supposed to ensure the rights of all sorts of adults.[5]

Rawls might respond to this by saying that at this stage he need not deal with every specific contingency. The principles of justice are very broad; the participants do not have to resolve every social problem before they can decide on principles. More specific rules will appear at a later stage in the development

of justice, at a time when they are framing laws rather than picking general principles. Rawls would be entitled to this reply if he himself had not already introduced so many specifics into the knowledge conditions for the original position. He denies a participant knowledge of his own strength, etc., so that he will abjure principles that favor any one group. Rawls therefore invites the question of whether, despite his precautions, any particular group will fail to receive equal consideration. Children form one such group because the participants must know they are adults.

IV

Thus far I have stressed the difficulties that stem from Rawls' knowledge restrictions in the original position. There are also problems arising from his rather strong requirements. The participants are supposed to know the "general facts about human society" (p. 137) about politics, economic theory, and "human psychology." This is rather a tall order even for the hypothetical original position. It seems particularly curious when the entry condition is considered along with it. "It must make no difference when one takes up this viewpoint, or who does so . . . the veil of ignorance is a key condition in meeting this requirement. It insures not only that the information available is relevant, but that it is at all times the same" (p. 139). The requirements on pp. 137 and 139 conflict. No one will be able to enter the original position if all sorts of abstract knowledge is required. While it may be possible at least in theory for the participants to suppress their knowledge of certain personal facts, it is a different matter to insist that they be furnished with a particular theory of society, its rules and its members. Even if we grant Rawls the claim that his views on economics, psychology and the rest are correct, there remains the problem of how these facts are to be given to the participants. And the entry condition requires that each of us be able to "adopt the perspective" of the participants, who are now made to be more knowledgeable than you or I. It is one thing to say that we can forget or ignore certain facts; it is quite another to say that we suddenly know facts we did not know before. We can convincingly pretend to be dumber or weaker than we are, but we cannot convincingly pretend to be smarter or stronger.

Indeed, when we examine Rawls' argument more closely, we see that the knowledge requirements do no work at all in the derivation. Of course, we have the intuition that accurate accounts of politics,

economics, and psychology would be desirable in choosing principles of justice. But since Rawls does not possess such an account, he cannot claim that possessing them would lead the participants to choose his principles. Indeed, our psychology may turn out to be such that the strains of commitment to Rawls' principles are too much to handle. At any rate, Rawls cannot now prove that this is not so; he cannot argue on the basis of information he (and we) do not have. The knowledge requirements serve merely as impressive window dressing.

Rawls might respond to all this by saying that he is not requiring that the participants know the truth about economics and the rest, but just that they use the best available theory. This is a much more reasonable demand, but it leaves Rawls with the problem of what happens with people from different centuries. People from the 18th century will have views on society that differ from those of people from the 20th or 22nd century. Even people from the same century might honestly disagree about, say, whether behaviorism or Freudianism offers the best approach to the human personality. And surely Rawls does not want 18th century justice as fairness to lead to different principles than does 20th century justice as fairness.

The knowledge conditions Rawls places on the original position make it unworkable. On the one hand participants are required to know facts they cannot know; on the other hand they are told to forget traits that will make their presence felt even if forgotten. Part of the trouble, we have seen, stems from the entry condition. It might be thought that some of these problems would disappear if the argument from hypothetical consent could somehow be retained without the entry condition. However, the entry condition cannot be eliminated. The following passages show that, for Rawls, the entry condition is paramount:

> To say that a certain conception of justice would be chosen in the original position is equivalent to saying that rational deliberation satisfying certain conditions and restrictions would reach a certain conclusion. If necessary, the argument to this result could be set out more formally. I shall however speak throughout in terms of the notion of the original position. (p. 138)

> In any case, it is important that the original position be interpreted so that one can at any time adopt its perspective. (p. 139)

The entry condition thus emerges not as a side issue but as the whole issue. The original position is intended to be a sort of shorthand symbol that stands for the entry condition. To suggest that the latter should be eliminated is to concede that Rawls' whole apparatus is unworkable.

V

The only way to salvage Rawls' system, then, is to alter the knowledge conditions. One alteration is suggested by the already-noted overlap of the formal constraints on the concept of justice and the veil of ignorance. Because these constraints forbid the choice of many tailor-made principles, we may lift the veil for such items as names, fingerprints and memories. Many controversial issues emerge here. Among them is the fact that a definite description can be recast as an indefinite description that happens to pick out one individual; another is the question of whether, say, the "maleness" of a particular man is uniquely his or just an instance of maleness indistinguishable from any other instance. A decision on these issues is not called for here. The point is that the situation is far more complicated that Rawls realizes.

A second alteration is this. Bias threatens because the participants are legislating for the society in which they themselves will live. This threat might vanish if the participants thought that they were legislating for others who were just like them in matters of psychology and distribution of talents. Let the participants believe that they are legislating for future human generations, but not for the one they live in now. In either case, the question of what the participants themselves get out of the principles they choose will not arise, at least directly. The participants' bias can be circumvented. It might be objected that even though they will not be personally affected, the participants will still tend to favor their sex or race or whatever group they belong to. However, even if they do have such a tendency it will be weaker than their bias toward their own positions. Another possible objection is that people choosing principles for others will not be as concerned as they would be if they were choosing for themselves. We shall have to assume that the participants meet in good

faith and that they choose principles carefully and conscientiously. But this is not too much to ask, since it is no more than Rawls assumed in the original position.

A more serious objection is that the shift to this altered original position weakens the argument from hypothetical consent. It now becomes a question of our adopting principles that we would pick for others rather than for ourselves. But this threat is not all that serious: on the assumption lately broached that the participants are acting in good faith, the participants would be willing to abide by the principles they selected. This even has the advantage of shoring up a weakness of the old original position: because of the veil of ignorance, the non-participant who is asked to adopt the principles of justice could say "If I had known then what I know now about myself, I would not have picked those principles." The participant in the new original position cannot say this. He is actually more apt to accept the principles he has chosen than is a participant in Rawls' original position. Once again, we find the knowledge restrictions causing trouble. Their elimination makes the hypothetical consent argument easier to take.

But while this shift to a more plausible original position strengthens Rawls' method, it has serious consequences for Rawls' theory. It seems unlikely that participants in the new original position will use the maximin strategy, a crucial premise in Rawls' "derivation" of his second principle of justice. Since the participants are no longer completely in the dark, they will probably employ some other strategy to guide their choice. Certainly they no longer have any reason to act as if they were infinitely averse to risk. Nor is this an accidental consequence of the altered original position I have sketched. The veil must be lifted if the entry condition is to be satisfiable, and as soon as it is lifted so is some element of potential risk.

Where does all this leave Rawls' system? The entry condition turned out to be the raison d'être of the original position. Since the knowledge conditions on the original position collide with the entry conditions, the former must be modified. One can make the veil of ignorance unnecessary by having the participants choose for others. This shift makes the hypothetical consent argument more plausible, but may lead to the rejection of maximin as the

strategy to guide the choice of principles. So in order to salvage the original position we must jettison those aspects of the theory--the knowledge conditions and their resulting strategy--that lead to Rawls' principles. The new original position may well lead to distinctly non-Rawlsian principles. To paraphrase a well-known explanation, we have to destroy the theory in order to save it.

University of Minnesota

NOTES

[1]Cambridge: Harvard University Press, 1971.

[2]See section 23, "The Formal Constraints of the Concept of Right."

[3]For an interesting discussion of hypothetical consent, see Michael Slote, "Desert, Consent, and Justice," Philosophy and Public Affairs, 1973.

[4]R.M. Hare, "Rawls' Theory of Justice," in Norman Daniels, ed., Reading Rawls (Basic Books: N.Y., 1975).

[5]Whether the participants will consider that they might become childlike through mental deterioration is a quite different matter. Degeneration may arise later, and anyway is not the norm. On the other hand, the participants know they are not young children. They may gamble that they will not become senile or brain-damaged; they know that at present they are competent adults. Of course, since Rawls assumes that the participants are infinitely averse to risk, he might argue that they would take the possibility of mental deterioration quite seriously. On the other hand, since it is normal to grow old, the participants will not choose principles that lead to mistreatment of old people--unless they are willing to gamble on a short life and a merry one.