Some Considerations About the Discovery of Principles of Justice

In *A Theory of Justice*, John Rawls claims to have presented a theory of justice in the tradition of the social contract theory. I shall not discuss all, or even many, aspects of his theory. I am not, for instance, concerned here with what Rawls thinks the principles of justice really are. I am concerned primarily with Rawls’s idea of the contract device as a tool for generating principles of justice.

One may read traditional social contract theories as attempts to answer the question: How can governments (or a certain kind of government) be justified? But Rawls is not initially occupied with this question. He tries to “generalize and carry to a higher order of abstraction the traditional theory of the social contract” (*TJ*, viii) in order to use the contract method for the examination of a more fundamental question: What are the (basic) principles of justice? (*TJ*, 11) We may identify those principles as those chosen by rational persons under certain conditions. The situation of choice—Rawls calls the “original position”. Although the original position is one specific version of the “initial situation”—a situation in which persons are to deliberate and to choose principles of justice—Rawls claims that it is the philosophically preferred one. (*TJ*, 17f, 118, and elsewhere) Two reasons motivate this claim. (1) We already have some fairly stable ideas about what will count as principles of justice. If we find that the deliberators in an

---

2 His “guiding aim” is to present an alternative to utilitarian and intuitionist conceptions of justice. But I shall not be concerned that issue.
3 It is a hypothetical situation, of course. I shall return to this point later.
initial situation would decide that, say, “Deaf persons are to be imprisoned for life” is to be a principle of justice, we may justifiably think that we should alter the description of the initial situation so that such a result would not appear. (TJ, 12, 120, 194) (2) In addition, we tend to think that some conditions are appropriate for deliberating about principles of justice and others are not. If, in an initial situation, one person is to announce what he holds to be principles of justice, and a second person is to electrocute the first, just in case the second person is not pleased with the suggested principles, then we may think that profitable deliberations have been put to a disadvantage. Rawls claims that there are certain preferred conditions which characterize the original position. Different constraints placed on the initial situation may yield different theories of justice, and the theory which centres on the original position is called justice as fairness: it is a theory about what would be chosen if a condition of fairness obtained in the initial situation.

In this paper I hope to make plausible the claim that the original position device presupposes a certain fundamental principle of justice.

I

Rawls is preoccupied with the idea of equality—or, I should say, with the fact of inequality. In “Justice as Fairness” he says that the “usual sense of justice” is “essentially the elimination of arbitrary distinctions and the establishment, within the structure of a practice, of a proper balance between competing claims.” (But what of

---

arbitrary equalities? Or are there such things?) In *A Theory of Justice* there appear many passages implying the (*prima facie*) undesirability of inequalities. (E.g., *TJ*, 7, 62.) Evidently we are to make a presumption in favour of equality. Inequalities, therefore, require justification. This leads Rawls to place considerable importance on the fact of so-called natural endowments. (*TJ*, 104) (Rawls does state, however—although not with much emphasis—that what should be said here is that the concept of desert does not even apply to such cases. (*TJ*, 102, 104) It is misleading, then, to say that one does not deserve one’s natural endowments.)

The idea of justice itself, Rawls says, always expresses a kind of equality—equality, that is, at least in the sense that a rule administered impartially implies “equality” to all persons or classes according to how that rule defines those classes. He might have been better off merely to state (as he finally does) that justice is the adherence to a rule. That, anyway, is its etymological meaning. The notion of equality arises by implication, if you are concerned to draw such an inference. Adherence to a principle or rule Rawls calls “formal justice”. But formal justice is not critical of what those principles are. That, apparently, is the business of “substantive justice”.¹⁶, ⁷ (*TJ*, 59)

---


¹⁷ There can be seen to be a close connection between the concept of justice and the concept of justification. Justification (as Rawls himself mentions at *TJ*, 107) is the inference of that which is to be justified on the basis of something else: it is, in fact, argumentation. Although a justification need not be deductive, it resembles
The idea of equality pervades *A Theory of Justice* not only because justice itself is supposed to express a kind of equality, but also because principles of justice “are needed for choosing among various social arrangements which determine [the] division of advantages and for underwriting an agreement on the proper distributive shares.” *(TJ*, 126; see also 7, 136, 171) The differences between persons according to their talents, intellectual abilities, strengths, and positions in society are mere facts about the world; they are neither just nor unjust. “What is just and unjust is the way that institutions deal with these facts.” *(TJ*, 102) The conclusion to be drawn is that “the primary subject of the principles of social justice is the basic structure of society, the arrangements of major social institutions into one scheme of cooperation.”8 *(TJ*, 54; see also 7f, 11, and else-

---

8 I am not sure why Rawls says “one scheme”. Perhaps it may be argued that institutions regulated by one set of principles of justice constitute a single scheme of cooperation. But why is there to be only one set of principles? To be sure, the hypothetical deliberators understand the advantages of cooperation; they understand that they must agree upon rules specifying the terms of their cooperation. But suppose they agree to certain rules for certain sorts of cooperative endeavours, and other rules for other sorts. Is this still “one scheme of cooperation”? And one set of principles of justice? If so, what would count as having more than one scheme of cooperation, and more than one set of principles?
More exactly, since, according to Rawls, the basic institutions of society determine persons’ rights, duties and shares of the benefits and burdens of social cooperation, “the principles of justice are regarded as formulating restrictions as to how practices may define positions and offices, and assign thereto powers and liabilities, rights and duties.” But why should there be such (or even any) restrictions? Why, that is, does the assignment of rights, duties and distributive shares create problems of justice?

But what does it matter that there may be conflicting claims? Why are principles for denying some claims and for sustaining others (or for reaching a compromise between competing claims) necessary? Well, if there are conflicting claims, not all can be satisfied. But in some cases of conflicts, the claimants will try to obtain satisfaction, and the most basic method of trying to obtain satisfaction—basic, that is, in the sense that it is always possible, even when all else fails—is force (or fraud). Rawls thus recognizes that it “is to avoid the appeal to force and cunning that the prin-

---

A similar question arises in response to traditional social contract theories. Rawls does not mention voluntary, private associations, as distinguished from society itself, but the principles of justice are not to apply, he says, to such arrangements. (TJ, 8)

9 “Justice as Fairness”, p. 316; also TJ, 6f.
principles of right and justice are accepted. Thus I assume that
to each according to his threat advantage is not a concep-
tion of justice.” (TJ, 134) And it is not, because it is not an
alternative to force.\textsuperscript{10} Apparently, then, the primordial mo-
tive is to avoid force. A similar point is made by H.L.A.
Hart, to whom Rawls refers the reader. But Hart, perhaps
like Hobbes, thinks that a fundamental motive is to avoid
death and bodily harm. If there were no rules against killing
and harming, then “what point could there be for beings
such as ourselves in having rules of \textit{any} other kind?”\textsuperscript{11} Of
course, remarks Hart, if things were otherwise—if, for ex-
ample, “men were to lose their vulnerability to each
other”—then the reasons for such a basic rule as “Thou
shalt not kill” might vanish. Yes, but what would not van-
ish would be the reasons for the basic rules expressed as a
prohibition, not against killing or harming, but more gener-
ally against coercion. Non-lethal force might occur not only
among people as we now know them, but also among in-
vulnerable persons: though I might not be able to kill or
harm such a being, I might nevertheless be able to block
his path, divert him to another path, or trick him.

\textsuperscript{10} Actually, the reason Rawls gives is stronger. It is not a conception
of justice, because “[i]t fails to establish an ordering in the required
sense, an ordering based on certain relevant aspects of persons
and their situation which are independent from their social posi-
tion, or their capacity to intimidate and coerce.” (TJ, 134) The rea-
son is stronger here because Rawls has already decided that prin-
ciples of justice must be chosen in a condition of fairness which
excludes the influences of one’s “social position”. But what has
the notion of social position to do with all of this? Apparently,
one’s capacity to intimidate may be linked to one’s social position.
But that is surely a weak link. And even were it not, what we ought
to be most concerned about is not simply intimidation, but rather
intimidation which is a threat of coercion.

If I am right that rules are necessary, in the first instance, in order to avoid, as much as possible, “appeals to Heaven” (to use Locke’s phrase), then have we not already hit upon a principle of justice? Certainly in the case of individual actions we now have such a principle. Let me name this the Libertarian Principle: Do not initiate force or fraud. But Rawls insists that the primary subject of justice is not individuals but rather the basic structure of society. Yet it seems that in merely setting the stage for considerations about justice, we have inadvertently stumbled onto a principle of justice, and justice for individuals at that.

It might be thought that Rawls’s idea of the concept of justice (as opposed to the various conceptions of justice) identified this basic desire (or need) to avoid appeals to force. But his distinction between the concept of justice and conceptions of justice does not accomplish that identification; in fact, his distinction makes that identification easy to miss:

Men disagree about which principles would define the basic terms of their association. Yet we may still say, despite this disagreement, that they each have a conception of justice. That is, they understand the need for, and they are prepared to affirm, a characteristic set of principles for assigning basic rights and duties and for determining what they take to be the proper distribution of the benefits and burdens of social cooperation. Thus it seems natural to think of the concept of justice as distinct from the various conceptions of justice and as being specified by the role which these

---

12 I say “Do not initiate force or fraud” in case we feel that the use of force is sometimes acceptable in response to force. It is not necessary, for the purpose of this paper, to include fraud in the prohibition, but I included it because we tend to place it on about the same level as (some kinds of) force.

13 *TJ*, 16, 110. See also “Justice as Fairness”, pp. 315ff.
different sets of principles, these different conceptions, have in common. (*TJ*, 5)

The claim here is that there is some problem, the solution for which is the establishment of principles of justice for institutions; and the realization that some such principles are needed is called the concept of justice. However, Rawls does not here clearly indicate what that problem is, and does not do so until later (*TJ*, 134) when he mentions what I shall call the “motive” for principles of justice, *viz.*, to avoid appeals to force. Had Rawls at the earlier point described this problem, it might have been more clearly recognized that the solution requires only some principles of other, but that a further argument is necessary in order to establish that those principles ought to be principles for *institutions*. My point is simply that the principles of justice which form the solution to the problem if appeals to force seem to be principles of justice for individuals.

II

Rawls has available two arguments for the claim that principles of justice for institutions must be prior to (or more fundamental than) principles of justice for individuals.\(^\text{14}\) The first is not found explicitly in *A Theory of Justice*, but was discussed sixteen years earlier in “Two Concepts of Rules”.\(^\text{15}\) In that article he distinguishes between the justification of rules and the justification of

\(^\text{14}\) There is, in addition, this statement: “The basic structure is the primary subject of justice because its effects are so profound and present from the start [of one’s life].” (*TJ*, 7) But that reason does not justify the conclusion. Perhaps Rawls ought to have said merely that the basic structure of society is an *important* subject of justice because its effects are so profound.

\(^\text{15}\) *Phil. Review*, 64 (1955) {pp. 3–32}. 
actions falling under a rule. But there are two concepts of rules: on the “summary” conception rules are generalizations from individual decisions or cases. On the “practice” conception rules define practices (institutions). On the former conception, rules are logically prior to the cases, since the classification of the individual cases can take place only on account of the rules which have characterized or defined them. Now, since, according to Rawls, persons’ rights, duties and distributive shares are defined in terms of practices (institutions), those practices are logically prior to the rights, duties and distributive shares. Thus, institutions must be the primary subject of justice: one criticises the assignments of rights and the distribution of shares on the basis of the institutions defining them; principles of justice, then, apply to those institutions themselves.

Now, it may be admitted that practices do indeed define rights, duties, and distributive shares. But does it sit well with our considered judgements\(^\text{16}\) that all rights, duties, and so on are defined by institutions?

Even if Rawls makes no use of the argument available from “Two Concepts of Rules” (and, after all, it is not explicitly used in *A Theory of Justice*), there remains another, perhaps more fundamental, consideration. Although he claims that his theory “generalizes and carries to a higher order of abstraction” the traditional social contract theories, it can be seen that the device of the original position represents a significant departure from (some) traditional theories in at least one respect. Traditionally, the

---

\(^{16}\) Considered judgements are “those judgments in which our moral capacities are most likely to be displayed without distortion. … Considered judgments are simply those rendered under circumstances favorable to the exercise of the sense of justice, and therefore in circumstances where the more common excuses and explanations for making a mistake do not obtain.” (*TJ*, 47–48)
parties to the social contract have the option of joining the society or not. If they had no such option, the conjunction of the concepts of the State of Nature and the social contract would make no sense. Indeed, the State of Nature is a state of no organized society. It is just the point of the traditional theories to explain how persons would find it reasonable to enter into a social contract. It is a further point of the traditional theories to argue that one kind of society—or social contract—would be more reasonable than another. Both Hobbes and Locke, for examples, make this distinction (even though it may not always be clear whether the two points are separated into two distinct contractual acts). It is true that the deliberators in Rawls’s original position are not, as in traditional theories, concerned simply with which society they shall organize. Rawls believes his theory to have abstracted the contractual method from that particular subject matter; he wishes to apply that method to principles of justice. But in fact his abstraction is not complete. The original position deliberators do not ask themselves “What are the principles of justice?”, but rather “What principles shall govern the institutions around which our cooperation shall be organized?” The principles chosen in response to this question are then named principles of justice.

We do believe, however, that there are, or can be, rights and wrongs, duties and obligations, justice and injustice, even without organized social relationships. But if that is so—if there could arise questions of justice even before there were no social practices—can the justice of such practices (institutions) be the primary subject of justice? One might indeed argue that without such institutions the prevention of injustice and the maintenance of justice would be a problematic matter. But that would not deny—in fact it would affirm—that principles of
justice have application there. What could be said, however, is that the principles of justice simply tell us what institutions there *ought* to be. But will deliberations about what institutions there ought to be be sufficient to announce what the principles are on the basis of which those institutions have been selected? From knowing that an action is unjust, it might follow that, if we wished to participate in social life, we would choose social institutions which would prevent or punish such actions. How reasonable is the reverse procedure? Can we discover what justice requires of individuals’ interactions, even in the absence of social practices, by first deliberating about what institutions we would choose, if we wished to participate in social life? Evidently it is exactly this reverse procedure that Rawls attempts to follow. Are the two procedures equivalent? It is not clear to me that they are.

Of course, Rawls wishes to limit his inquiry to “social justice”, which, he says, concerns the basic structure (i.e. the major institutions) of society. (TJ, 7) Does this give him sufficient reason to restrict the original position in the way he does, i.e. so that institutions are the primary subject of deliberation? I think not, because, as I have argued, from a less restrictive view, the principles of justice for individuals must be decided upon first (or, at best, the deliberators in the original position, even though they are supposed to choose principles for institutions, would understand that the principles for individuals are the primary subject of justice and would restructure their inquiries accordingly).

In discussing principles of justice for individuals, Rawls says that “obligations presuppose principles for social forms”. (TJ, 10) I believe, in the light of my comments so far, that it can be the other way around. Here is Rawls again: “that principles for institutions are chosen first shows the social nature of the virtue of justice....” (TJ,
But on the contrary it shows no such thing, or, rather, it would, but only if the deliberators chose to choose principles for institutions first and principles for individuals next. I have argued that just the reverse would take place. On the other hand, if the original position is structured in such a way as to make it impossible that choosing principles for individuals could be other than a secondary matter, then it is misleading to say that the social nature of justice has been shown. No. Perhaps it has been stipulated, but not shown. And such a stipulation would not fit well with our (or at least my) considered judgements.

If I am right—if principles of justice for individuals are the primary subject of justice—then the bare motive for wishing to discover principles of justice gives rise to one such principle. If principles of justice are sought in order to find alternatives to the settlement of conflicts by force, then one such principle is immediately available, namely, the Libertarian Principle. And this principle makes sense even prior to questions about institutions.

III

What is the purpose of the requirement of unanimity in the original position? In the contract theories of both Hobbes and Locke, unanimity occurs only by exclusion—it is an *ad hoc* characterization of the move to a commonwealth. That is, it can be said that after a contract has been made, all parties to it have consented; hence, consent to the contract has been unanimous among the contractees. That, of course, is analytic; it could not be otherwise without altering the meaning of contract. But both Hobbes and Locke allow for the possibility of there being persons in the State of Nature who do not enter into
the contract. So while contract, by definition, requires
unanimity among all the contractees, it does not require
unanimity among all the persons originally in the State of
Nature. Out of the State of Nature a commonwealth—
indeed, any number of commonwealths—may emerge
without requiring unanimity among all natives of the State
of Nature. In addition, Locke, who is perhaps the clearest
of the contract theorists on this point, announces that no
one is a member of the resulting commonwealth(s) who
does not consent to be a member. But why not? Perhaps
Locke thought it a violation of a fundamental principle to
take action toward a person without that person’s consent.
If so, then that principle itself might do as a principle of
justice. In fact, it is a form of the Libertarian Principle.

Rawls’s version of the contract method is not merely
an abstraction from traditional theories; it distorts the
contract method by its peculiar use of the unanimity
requirement. Deliberators in the original position are to
choose principles, but unanimity of all persons is required if
there are to be any principles. If that unanimity is not to be
had, then there are to be no principles, and the result is
what Rawls calls general egoism, according to which
“everyone is permitted to advance his interests as he
pleases”. (TJ, 124)

Let the traditional unanimity requirement—a
consensus among contractees only—be called contractual
unanimity, and let the unanimity which Rawls requires—a
consensus among all persons initially eligible to express an
opinion—be named unrestricted unanimity. Why does Rawls
insist on unrestricted unanimity? Well, traditional,

17 There may be situations in which a tacit consent is recognized as
obliging a person to obey the laws of the commonwealth. Even
then, however, the tacit consent does not make a person a
member of the society.
contractual unanimity establishes a contract, an agreement—a convention. The concepts of truth and falsity do not apply, because an agreement to form an association is neither true nor false. But it would not fit our considered judgements to suppose that principles of justice are to be mere conventions; hence, Rawls insists on unrestricted unanimity, and his contract theory, unlike the traditional theories, is to be taken as a *method of discovery*—a discovery of, if not what is true, then at least what is reasonable.

Both Rawls’s theory and the traditional theories are theories about hypothetical situations. But of what value are deductions made from hypothetical constructions? In science, a theory may be proposed as a hypothetical construction offered for testing. Deductions made from the hypothesis, with the aid of background beliefs, may be used as predictions—as guides for the collection of data. If data are collected which accord with the hypothesis, there may be a tendency to feel confident in the truth of that hypothesis. That is to say, we have a tendency to rely on the hypothesis to predict accurately the character of further data. And so it is with the hypothetical original position. We are to deduce certain principles to be used as principles of justice. If these principles tend to coincide with our considered judgements, then we may have a tendency to rely on other implications from them concerning matters for which we may not have an antecedent or stable considered judgement.18 Once the hypothetical original position is finally fixed, then “to say that a certain conception of justice would be chosen in the original

---

18 The result of changing the hypothesis so as to yield principles in accordance with our considered judgements and changing our less settled considered judgements to match the deductions from the hypothesis, Rawls calls “reflective equilibrium”. (*TJ*, 20f, 48f)
position is equivalent to saying that rational deliberation satisfying certain conditions and restrictions would reach a certain conclusion.” (TJ, 138) That is, the choice of principles of justice is an objective matter, understandable by means of one’s deductive powers, and so, to understand the reasoning is to understand that “if a conception of justice would be agreed to in the original position, its principles are the right ones to apply” (TJ, 169f), which is to say that “our [present] social situation is just if it is such that by this sequence of hypothetical agreements we would have contracted into the general system of rules which defines it.” (TJ, 13)

Let me approach the matter from a different angle. What does it matter whether in the original position we would consent to certain principles? Why does Rawls attempt to persuade us now about something we would do? Rawls’s aim is not merely to convince us about what, in some hypothetical situation, we would do; the hypothetical situation is described in such a manner as to bring about our present, actual agreement on principles of justice. It is our real consent to certain principles that is finally desired. So it is in the case of any argument: by means of certain considerations, and by following certain standards of fairness and rationality, an argument is advanced in hopes of winning assent from an audience. Actual unanimity is a test of a theory which predicts hypothetical unanimity.

But if real consent is the value ultimately sought, what is the appropriate response to persons who withhold their real consent? Suppose, for example, we witness a situation in which $A$ acts toward $B$ in a manner which we think to be unjust—in a manner, that is, which we think neither $A$ nor $B$ (nor, of course, we ourselves) would agree to in the original position. And suppose we announce this belief to $A$ and $B$, both of whom agree with our assessment
of the original position. But suppose, finally, that both \( A \) and \( B \) say: “So what? In the original position we would decide one way. Big deal. This is not the original position. What’s more, there is really nothing unjust in the case of our present interaction.” What responses can we appropriately make to this? If we follow *A Theory of Justice* we might still label the situation between \( A \) and \( B \) unjust. (We might even believe that we were justified in taking interfering action.) Perhaps, in addition, we would say that \( A \) and \( B \) were blockheads who did not understand the idea of the original position after all, or who were simply not rational beings. On the other hand, we might wish to revise the nature of the original position, or to reconsider whether principles chosen there would really be principles of justice.

Well? Which shall it be? My considered judgements on this matter tell me to make a *presumption* in favour of what \( A \) and \( B \) really say, as opposed to what I or anybody else think they *would* say under different circumstances.\(^{19}\)

Let us assume for the moment that coercion can be an appropriate response to unjust situations. If we think that \( A \) acted (or is acting) unjustly toward \( B \) and if we understand that this kind of injustice may call for interfering action, then we may feel justified in interfering with \( A \)'s actions. So we interfere with \( A \). But, as before, suppose that both \( A \) and \( B \) complain about our interference. The basic motive for justice, remember, is to provide an alternative to appeals to force. (Force, in the sense relevant here, might be more specifically described as force without the consent of the persons involved. That is, force used *with* consent does not give rise to the same motive to seek alternatives.) But appeals to force arise in situations where there are conflicts of interests. And ought

it not at the very least be *presumed* that there is no conflict of interests where the persons involved actually claim there is none? In the above example, interfering with \( A \) (and/or with \( B \)) did not have the consent of either \( A \) or \( B \), never mind what we or they or anybody else would consent to under different circumstances. Such interfering actions are actions against the real, present wishes of the persons involved, and so even if our interference is an attempt to do justice—or, say, to act according to some precept of rectificatory justice—we contradict the initial motive for justice (and break the Libertarian Principle).

There may be situations in which we might think that interfering action is required even without the consent of the person(s) involved. But what is the force of the “without” here? There are at least two possibilities. (a) Consent is not to be had because the person is simply unable either to give or to withhold consent—an unconscious person, for example. In such a situation it might make sense to *presume* that the person would give consent to interfering actions if he were able to. But that is not to say that he would give his consent (to the principle which would justify our interference) if he were in some specially restricted condition of fairness. Rather, it is to say that we believe that in these present circumstances, changed only so as to allow for his being able to give or to refuse consent, he would consent. If we allow for this presumption, we weaken the Libertarian Principle somewhat. (b) On the other hand, “without consent” might be taken to mean “against the refusal to consent”, and in this case a substantial weakening of the Libertarian Principle.

---

20 If a person is able to give or to refuse consent, but remains silent, then there may in some circumstances be reasons for calling this a “tacit consent”. I shall side-step this issue here, noting only that the notion of tacit consent can easily be abused.
Principle occurs. The result might be called the *Liberal Principle*: Do not act against dissenting persons, *unless there are good reasons to the contrary*. The rider here is of sufficient consequence to generate a great deal of further puzzles. Perhaps, after all, *A Theory of Justice* can be seen as an attempt to describe and to justify some such “good reasons to the contrary”.

David B. Suits, Rochester Institute of Technology