ON SOCIAL JUSTICE

Charles Taylor

Robert Paul Wolff, *Understanding Rawls: A Reconstruction and Critique of A Theory of Justice*, Princeton: Princeton University Press, 1977, $13.50 cloth, $3.95 paper.

There have been literally dozens of articles, and some books, written in comment on John Rawls' *A Theory of Justice* since it appeared in 1971. A great many of these have attempted to demonstrate how Rawls' famous derivation of the two principles of justice from the original position through game-theoretical reasoning does not really work.

Robert Paul Wolff also tries to show that this derivation breaks down, but his book attempts much more. It attempts to justify its title. Wolff's aim is to make clear the "basic idea" or the "core insight" which informs Rawls' theory through its many versions, from the early articles in the 1950s through to the mature statement of his position in *A Theory of Justice.*

In Wolff's view, Rawls' core insight offers a way out of an impasse in which many sensitive philosophically inclined people find themselves in the Anglo-Saxon world once they begin to reflect on the bases of their ethical position. They find themselves torn between utilitarianism on the one hand, and on the other some view which will make sense of their strongly felt moral intuitions concerning the unconditional nature of the right, and, in particular, the inviolability of the human person. For some this alternate view might take the form of intuitionism, but even those who are quite unattracted to intuitionism as an ethical theory often seek some way of grounding these moral intuitions.

What strengthens the appeal of utilitarianism in this philosophical culture is that it offers a clear method of reasoning about ethical matters which fits well with our paradigms about reasoning, viz., calculation; and, moreover, calculation about an unchallengeable, unmystical and thoroughly empirical definition of the human good, viz., happiness. What is unappealing about utilitarianism, apart from its general insensitivity and crassness, is that it does not seem to allow for an unconditional right and wrong. What is right depends on what is good, on what will produce the greatest quantity of good in any situation, i.e., what tends to the greatest happiness of the greatest number. This seems to permit us to reason in certain (admittedly boundary) situations about the admissibility of sacrificing some innocent person, or grossly neglecting his rights, in order to bring about the happiness of many others. But this, of course, contradicts our deeply held moral intuitions about the inviolability of the person.

The two sides of the liberal outlook thus enter into conflict with each other. On the one hand, this outlook starts from the individual and his rights as the ground for all ethical and political reasoning; the rest of the universe, including the political and social structures linking individuals, is to be conceived only as means to the securing of the rights and well-being of individual human beings. On the other hand, this very invitation to calculate the effects of nature and social structure *as means* tends to carry forward under the weight of its own intellectual justification until the fate of individuals themselves is part of the calculation. Liberalism is torn between its Lockean and Hobbesian sides; or, in terms of later reference points, between its utilitarian and Kantian allegiances.

Rawls' basic idea may be understood as providing a way out of this dilemma. Rawls, as Wolff argues, draws on another strand of the modern liberal outlook, contract theory, to produce a justification of unconditional right and the inviolability of the individual (in the form of a theory of justice which ensures that one cannot be sacrificed for others) by means of a rigorous argument of instrumental reason which involves attributing no controversial, substantive goals to the contractors, but which only assumes that they are interested in their own happiness. Consequently, the principles of justice, and the acceptance of inviolability that they incorporate, may be envisaged as the necessary outcome of an attempt to arrive at unanimous agreement about the rules governing their association on the part of self-interested individuals in certain defined conditions. If this argument holds, we would then have justified our most cherished intuitions about the right, but by means of a reasoning process every bit as rigorous and "tough-minded" as the utilitarians', and which, moreover, involves no questionable initial assumptions concerning what men necessarily seek or ought to seek. To take this same point from another angle, we should have proved some very substantive conclusions about how men ought to treat each other starting from some minimal, and purely formal assumptions: that men are self-interested (they have *some* goals, but we do not know which), that they must reach binding agreement on rules, that none can dominate the others, etc.

Wolff's purpose in *Understanding Rawls* is to take this core idea, and to trace its development through the different stages of Rawls' position, from the early "Justice as Fairness" article (*Philosophical Review*, 1958) to its mature statement in *A Theory of Justice*, accounting for the changes it has undergone in terms of the difficulties that it has encountered at each stage. This occupies Part Two of his work (Part One sets out the basic idea itself). In Part Three he discusses the relation of Rawls to Kant. In Part Four, he offers a wide-ranging criticism of the mature statement of the argument, and attempts to show that the derivation of the two principles in *A Theory of Justice* breaks down. Part Five provides some general reflections by way of conclusion on the inap-

propriateness of attempting to resolve questions about social justice in such an abstract fashion as that to which the apparatus of the social contract and game-theoretical calculation condemns us.

There is something of real moment in Wolff's notion that Rawls' theory should be viewed as a resolution of a dilemma, or at least of a tension, between the pull of utilitarianism and certain strong moral intuitions. This does much to account for the extraordinary popularity of Rawls' work, and the intense interest that it has generated in the philosophical world. Even though most of those who write about Rawls do so in order to refute him, they are drawn by the intrinsic interest of what he attempts, which is to open out this area of our strongest moral convictions to the same rigorous, calculative mode of reasoning which has achieved such prestige in other, less humanly and emotionally central, areas. In the intellectual culture of Anglo-American philosophy, where this mathematically-modelled argument enjoys such (I think irrational) prestige, it is a tremendous achievement when someone allows us to discuss something really humanly and philosophically important in this canonical form — the only one in which we can be sure of saying something philosophically valid.

From this point of view, Rawls' achievement can be seen as that of bringing together a certain content and a certain form of argument. It resolves the problem of those who might have felt a nagging, half-admitted worry that their mode of philosophizing was keeping them from addressing important questions (as the critics of analytic philosophy have always insisted). But Wolff's critique goes further. He contends that Rawls' core idea is meant to resolve the dilemma in moral philosophy outlined above, the tension between the utilitarian and the Kantian in the contemporary philosopher. I think there is a great deal in this, too. But there is also something very puzzling when one tries to clarify what this means.

Wolff himself is puzzled, for at the end of Part Four, he has a section entitled "The Logical Status of Rawls' Argument", where he offers three possible accounts, incompatible with each other, of what exactly Rawls might be trying to prove about his two principles by their derivation from the original position. And there is, indeed, a great mystery surrounding this question, which makes it very difficult to say exactly in what way Rawls can be seen as resolving the tension between the two sides of the liberal outlook.

Perhaps, at this point, I might share my own bafflement with the reader, and then refer him to Wolff's instructive discussion of this issue, and through that back to the text of *A Theory of Justice.* Wolff mentions three possible readings of what Rawls attempts to prove. I should like to single out two, which are close to his first two.

One might think that the derivation of the two principles from the original position was itself a proof of their validity. How might this be? One way might

be as an example of what Rawls calls pure procedural justice. We speak of pure procedural justice where the fairness of a distribution, for instance, resides in its having issued from a certain procedure. If we play a fair game of poker, and I lose my shirt to you, it is justly yours, in virtue of the way the game has actually gone.

But, as Rawls points out, it is essential to pure procedural justice that we actually play out the procedure. You could not walk off with my shirt before the game, and justify yourself on the grounds that this is a *possible* outcome of a poker game, or even that this is the *inescapable* outcome of a poker game (given my well-known combination of stupidity and rashness) between us. But now Rawls' contract is not something that we actually play out as contractors; it is an imagined predicament about which we are engaged in demonstrating the best strategy it dictates to us. So we cannot understand Rawls as intending the derivation as a proof of the two principles by pure procedural justice.

Another way exists of viewing the derivation as a proof of the validity of the principles of justice. We could envisage it as a claim that a rational agent, in the sense of an agent of instrumental reason, was committed to these principles as his best strategy (on pain thus of irrationality), once he accepted that he had to enter into *some* binding system of rules with others. This seems to be Wolff's view of Rawls' original intentions. Thus, discussing Rawls' early position, he states: "Rawls would, if he could prove his theorem, be in a position to say to a reader:

> If you are a rationally self-interested agent, and if you are to have a morality at all, then you must acknowledge as binding upon you the moral principle I shall enunciate." (p. 17)

If this theorem could be demonstrated successfully, one would have solved the tension between utilitarianism and our intuitions about right; for from a basis no richer than that of utilitarianism, viz., the self-interested individual, plus the constraint that we must come to some binding agreement on the rules which are to hold among us (and surely we must all accept *this*, unless we are willing to live as hermits in the Mackenzie delta), we should have derived valuable moral notions.

However, whether Wolff is correct or not about Rawls' original intentions, this cannot be the status of the argument proposed in *A Theory of Justice*. This is sufficiently obvious from Rawls' discussion of reflective equilibrium, his explicit discussion of the possible need to adjust our definition of the original situation in order to derive principles that meet our intuitions, his invocation of Kant, and much else.

But, in addition, one can argue that it is just not possible to conceive of this derivation as a proof that the principles are valid. Even if we waive all of the objections that have been made to the derivation in *A Theory of Justice*, there are two basic reasons which rule it out as a validity proof, one touching the way things are, the other the nature of moral obligation.

The first reason is that the strategy of adopting the two principles is only shown to be the best for contractors in the extremely counter-factual predicament of the original position, where none has a special bargaining leverage to impose on the others his solution, and where, moreover, all are ignorant of their goals, talents, desires, etc. But in any real life contracting situation, we all know something of our goals, and of our bargaining counters, and it is rational to use these to the hilt. The two principles are far from being the rational policy for a self-interested contractor as such, but only at best in those empirically unrealizable conditions that Rawls lays down.

The second reason is that even if accepting the two principles did turn out to be the best strategy for rational agents as such, it would be just that: the best strategy of instrumental reason. It would still not have the status of a moral obligation, laying a higher claim on us than the realization of self-interest. But this sense of a higher claim is an integral part of the moral intuition we are trying to recapture. And this cannot be accomplished by an argument about the best strategy of instrumental reason. The gap here is the one Kant tried to mark by his distinction of categorical and hypothetical imperatives.

But if the derivation of the principles is not a proof of their validity then what is it? I should like to suggest that Rawls sees it as a method for defining justice. I want to distinguish a method of defining justice from a statement of what justice is. Perhaps, we could clarify this distinction by noting that there are two ways we could answer the question: What acts are right? or the question: How can I tell what acts are right? One would be to give a characterization of right actions which made clear in some way what it is that makes an action right, or as we might put it, that in virtue of which actions are right. We might reply for instance that actions are right which fulfil our nature as rational animals, or that tend to produce the greatest happiness of the greatest number. In both these cases we would be replying by providing the underlying ground that makes actions right (that is, our very controversial view of this ground). We are not only telling our questioner how to identify right acts from wrong acts, but we are also telling him why these acts are right or wrong.

But we might also reply in this vein. If you want to be able to tell right from wrong, then follow this rule: do unto others as you would be done by. Here there is no claim that what makes an act a right/wrong one is that we would want/not want it to be done to us. Rather the answer to *this* question might be: *it is* according to the will of God to treat your fellow creatures with

benevolence; or, we are bound by our common nature to do good and not harm to each other. The claim is only that the reflection: Would I want this done to me? provides an excellent (perhaps even infallible) method of discovering in any case what is right. What makes this right is something different.

The claim underlying such a criterion, of course, must be that there is some systematic connection between the criterion and what makes things right. We can see the grounds of this systematic connection in the case of the golden rule. If the basis of right and wrong is that we are called on by God or by nature to treat our fellow men with benevolence; and if we can assume that we are all roughly the same in our make-up; then a good criterion for whether I should do A to X (i.e., whether it is a benevolent act) is my willingness to have A done to me. We thus have a systematic connection between the grounds of right and our criterion; but they are not the same.

The golden rule is, as I noted above, a method of defining the right. Another famous such example, this time in the history of political theory, is Kant's use of the social contract idea, which is the direct ancestor to Rawls'. Kant suggests that we use the hypothetical test of unanimous agreement to ascertain whether laws are just. But this does not mean that something would be made just by the fact of unanimous agreement. What makes something just is that it can be willed as universal. There is a systematic connection between what can be willed as a universal law and what self-interested persons with varying goals will actually agree to unanimously; for in order to reach unanimous agreement, they would have to abstract from particular interests, and seek only what was in the interest of everybody. But in doing so they would have to detach themselves from the same particular goals which the moral person is asked to set aside as the motive of his/her actions. Thus we can expect a congruence between the unanimous compacts of even bad persons and the moral will of the good person.

I would argue that Rawls is proposing something of this sort in his derivation of the principles of justice. As a method of calculating what is just, it is very similar to the Kantian contract notion on which it is based; and it is meant to work as a criterion through the same kind of systematic connection (which Rawls discusses in section 40 of *A Theory of Justice*). Consequently, there is no implied claim that the derivation provides us with the grounds for just acts being just, much less, therefore, with a proof of their being just (which would have to lay clear the grounds).

Here I also take issue with Wolff. I do not agree that we can view the derivation as offering what he calls a "rational reconstruction" of our moral convictions (p. 181). For such reconstructions, which derive our multiform moral convictions from some small set of general principles, must also claim to lay bare to some extent the grounds of right; while what I have called a method of defin-

ing the right makes no such claim whatever.

Rawls' derivation as I see it makes no such claim. Rawls' notion of the grounds of right seems to be similar to Kant's (on a plausible interpretation of Kant). Embracing the principles of justice "expresses our nature as free and equal rational persons" (*A Theory of Justice*, p. 256). The ground of the right is that we are called on to live up to our status as rational agents. This requires that we judge universally, abstracting from our particular goals. There is thus a systematic connection between this process of moral abstracting, and that which self-interested contractors would be forced to in the original position. But the fact that the principles are a good strategy for self-interested subjects in certain conditions has nothing to do with what makes them principles of justice.

What then is the value of the derivation? Why not just argue for the principles directly out of their grounds, which in Rawls' case seem to be Kantian-Humboldtian in nature? (cf. particularly, section 79 of *A Theory of Justice*) The justification would be that of all methods of definition. What underlies the popularity of the golden rule? The fact that it provides an easily available identification of the right, even for those who find it difficult to reason from the content of God's will or from the demands of a common nature, or even from the requirements of general benevolence.

One might in a parallel way make a claim for the derivation, that it allows us to go further in an exact and fine-tuned definition of the principles of justice than we could achieve if we argued straight from the grounds of justice. Precisely because we can use game-theoretical reasoning, we can arrive at such finely-nuanced definitions of the general principle of social equality as the difference principle. Once such principles are derived in the game-theoretical argument, they are recognized by our intuitions and informed by our understanding of the grounds of justice as indeed principles of justice. In the course of this reflection, our intuitions have thus been changed (made more precise), and we will have reached what Rawls calls reflective equilibrium.

This, I submit, is one way of understanding the logical status of Rawls' contract argument. It makes sense of this argument, and gives it a justification which does not involve our making untenable claims about the status of validity proof on its behalf.

But if this makes sense of the argument, then why did I confess to bafflement above? Because it is not at all clear that this is the sense that Rawls makes of it. Much of Rawls' commentary in *A Theory of Justice*, including, for instance, the opening remarks about the value of contract theory, seem to give it a more exalted status than simply (as what I have called) a method of defining the just.

But there is more. If Rawls' contract has the same status as Kant's and as the golden rule, then it must surely be clear by now that it has failed in its purpose.

The derivation of the two principles is, even if valid, so complex and involves so much artifice that all hope must be abandoned of its serving as a vehicle of discovery of more rigorous, fine-grained definitions of the principles of justice. Rawls has a very interesting position on justice, derived partly from Kant and von Humboldt, which is being obscured by all the arid churning of the academic game-theoretical mill.

But I shall stop here, just short of saying something unbearably paradoxical: that Rawls' position might be improved by sloughing off what has made it the philosophical *succès d'estime* of the 1970s.

Perhaps one day, Rawls in his replies to critics shall take up the question of the logical status of the derivation, or what its strategic role is in the whole argument about justice. Should he do so, I would hope that he would address himself to this book. For Wolff, in trying to lay bare the core insight behind Rawls' developing position, has raised this issue with unusual focus and clarity. He brings together a grasp of Rawls' strategic goals with a detailed understanding of his arguments to produce an uncommonly interesting commentary on issues of pressing philosophical and moral significance.

All Souls College
Oxford, England