The Original Position

Ronald Dworkin†

I

I trust that it is not necessary to describe John Rawls's famous idea of the original position in any great detail. It imagines a group of men and women who come together to form a social contract. Thus far it resembles the imaginary congresses of the classical social contract theories. The original position differs, however, from these theories in its description of the parties. They are men and women with ordinary tastes, talents, ambitions, and convictions, but each is temporarily ignorant of these features of his own personality, and must agree upon a contract before his self-awareness returns.

Rawls tries to show that if these men and women are rational, and act only in their own self-interest, they will choose his two principles of justice. These provide, roughly, that every person must have the largest political liberty compatible with a like liberty for all, and that inequalities in power, wealth, income, and other resources must not exist except in so far as they work to the absolute benefit of the worst-off members of society. Many of Rawls's critics disagree that men and women in the original position would inevitably choose these two principles. The principles are conservative, and the critics believe they would be chosen only by men who were conservative by temperament, and not by men who were natural gamblers. I do not think this criticism is well-taken, but in this essay, at least, I mean to ignore the point. I am interested in a different issue.

Suppose that the critics are wrong, and that men and women in the original position would in fact choose Rawls's two principles as being in their own best interest. Rawls seems to think that that fact would provide an argument in favor of these two principles as a standard of justice against which to test actual political institutions. But it is not immediately plain why this should be so.

If a group contracted in advance that disputes amongst them would

† Professor of Jurisprudence and Fellow of University College, Oxford University. I have benefited from discussion of a draft of this paper in a seminar in Oxford in the fall of 1972, and from discussions with H.L.A. Hart and Thomas Nagel. Nagel raises some of the issues discussed in the first part of this paper in his review of Rawls's book, *Rawls on Justice*, 82 PHILOSOPHICAL REVIEW 220 (1973).

be settled in a particular way, the fact of that contract would be a powerful argument that such disputes should be settled in that way when they do arise. The contract would be an argument in itself, independent of the force of the reasons that might have led different people to enter the contract. Ordinarily, for example, each of the parties supposes that a contract he signs is in his own interest; but if someone has made a mistake in calculating his self-interest, the fact that he did contract is a strong reason for the fairness of holding him nevertheless to the bargain.

Rawls does not suppose that any group ever entered into a social contract of the sort he describes. He argues only that if a group of rational men did find themselves in the predicament of the original position, they would contract for the two principles. His contract is hypothetical, and hypothetical contracts do not supply an independent argument for the fairness of enforcing their terms. A hypothetical contract is not simply a pale form of an actual contract; it is no contract at all.

If, for example, I am playing a game, it may be that I would have agreed to any number of ground rules if I had been asked in advance of play. It does not follow that these rules may be enforced against me if I have not, in fact, agreed to them. There must be reasons, of course, why I would have agreed if asked in advance, and these may also be reasons why it is fair to enforce these rules against me even if I have not agreed. But my hypothetical agreement does not count as a reason, independent of these other reasons, for enforcing the rules against me, as my actual agreement would have.

Suppose that you and I are playing poker and we find, in the middle of a hand, that the deck is one card short. You suggest that we throw the hand in, but I refuse because I know I am going to win and I want the money in the pot. You might say that I would certainly have agreed to that procedure had the possibility of the deck being short been raised in advance. But your point is not that I am somehow committed to throwing the hand in by an agreement I never made. Rather you use the device of a hypothetical agreement to make a point that might have been made without that device, which is that the solution recommended is so obviously fair and sensible that only someone with an immediate contrary interest could disagree. Your main argument is that your solution is fair and sensible, and the fact that I would have chosen it myself adds nothing of substance to that argument. If I am able to meet the main argument nothing remains, rising out of your claim that I would have agreed, to be answered or excused.

In some circumstances, moreover, the fact that I would have agreed

does not even suggest an independent argument of this character. Everything depends on your reasons for supposing that I would have agreed. Suppose you say that I would have agreed, if you had brought up the point and insisted on your solution, because I very much wanted to play and would have given in rather than miss my chance. I might concede that I would have agreed for that reason, and then add that I am lucky that you did not raise the point. The fact that I would have agreed if you had insisted neither adds nor suggests any argument why I should agree now. The point is not that it would have been unfair of you to insist on your proposal as a condition of playing; indeed, it would not have been. If you had held out for you proposal, and I had agreed, I could not say that my agreement was in any way nullified or called into question because of duress. But if I had not in fact agreed, the fact that I would have in itself mean nothing.

I do not mean that it is never relevant, in deciding whether an act affecting someone is fair, that he would have consented if asked. If a doctor finds a man unconscious and bleeding, for example, it might be important for him to ask whether the man would consent to a transfusion if he were conscious. If there is every reason to think that he would, that fact is important in justifying the transfusion if the patient later, perhaps because he has undergone a religious conversion, condemns the doctor for having proceeded. But this sort of case is beside the present point, because the patient's hypothetical agreement shows that his will was inclined towards the decision at the time and in the circumstances that the decision was taken. He has lost nothing by not being consulted at the appropriate time, because he would have consented if he had been. The original position argument is very different. If we take it to argue for the fairness of applying the two principles we must take it to argue that because a man would have consented to certain principles if asked in advance, it is fair to apply those principles to him later, under different circumstances, when he does not consent.

But that is a bad argument. Suppose I did not know the value of my painting on Monday; if you had offered me $100 for it then I would have accepted. On Tuesday I discovered it was valuable. You cannot argue that it would be fair for the courts to make me sell it to you for $100 on Wednesday. It may be my good fortune that you did not ask me on Monday, but that does not justify coercion against me later.

We must therefore treat the argument from the original position as we treat your argument in the poker game; it must be a device for calling attention to some independent argument for the fairness of the two principles—an argument that does not rest on the false premise that a hypothetical contract has some pale binding force. What other

argument is available? One might say that the original position shows that the two principles are in the best interests of every member of any political community, and that it is fair to govern in accordance with them for that reason. It is true that if the two principles could be shown to be in everyone's interest, that would be a sound argument for their fairness, but it is hard to see how the original position can be used to show that they are.

We must be careful to distinguish two senses in which something might be said to be in my interest. It is in my *antecedent* interest to make a bet on a horse that, all things considered, offers the best odds, even if, in the event, the horse loses. It is in my *actual* interest to bet on the horse that wins, even if the bet was, at the time I made it, a silly one. If the original position furnishes an argument that it is in everyone's interest to accept the two principles over other possible bases for a constitution, it must be an argument that uses the idea of antecedent and not actual interest. It is not in the actual best interest of everyone to choose the two principles, because when the veil of ignorance is lifted some will discover that they would have been better off if some other principle, like the principle of average utility, had been chosen.

A judgment of antecedent interest depends upon the circumstances under which the judgment is made, and, in particular, upon the knowledge available to the man making the judgment. It might be in my antecedent interest to bet on a certain horse at given odds before the starting gun, but not, at least at the same odds, after he has stumbled on the first turn. The fact, therefore, that a particular choice is in my interest at a particular time, under conditions of great uncertainty, is not a good argument for the fairness of enforcing that choice against me later under conditions of much greater knowledge. But that is what, on this interpretation, the original position argument suggests, because it seeks to justify the contemporary use of the two principles on the supposition that, under conditions very different from present conditions, it would be in the antecedent interest of everyone to agree to them. If I have bought a ticket on a longshot it might be in my antecedent interest, before the race, to sell the ticket to you for twice what I paid; it does not follow that it is fair for you to take it from me for that sum when the longshot is about to win.

Someone might now say that I have misunderstood the point of the special conditions of uncertainty in the original position. The parties are made ignorant of their special resources and talents to prevent them from bargaining for principles that are inherently unfair because they favor some collection of resources and talents over others. If the man in the original position does not know his special interests, he

cannot negotiate to favor them. In that case, it might be said, the uncertainty of the original position does not vitiate the argument from antecedent interest as I have suggested, but only limits the range within which self-interest might operate. The argument shows that the two principles are in everyone's interest once obviously unfair principles are removed from consideration by the device of uncertainty. Since the only additional knowledge contemporary men and women have over men and women in the original position is knowledge that they ought not to rely upon in choosing principles of justice, their antecedent interest is, so far as it is relevant, the same, and if that is so the original position argument does offer a good argument for applying the two principles to contemporary politics.

But surely this confuses the argument that Rawls makes with a different argument that he might have made. Suppose his men and women had full knowledge of their own talents and tastes, but had to reach agreement under conditions that ruled out, simply by stipulation, obviously unfair principles like those providing special advantage for named individuals. If Rawls could show that, once such obviously unfair principles had been set aside, it would be in the interest of everyone to settle for his two principles, that would indeed count as an argument for the two principles. My point—that the antecedent self-interest of men in the original position is different from that of contemporary men—would no longer hold because both groups of men would then have the same knowledge about themselves, and be subject to the same moral restrictions against choosing obviously unfair principles.

Rawls's actual argument is quite different, however. The ignorance in which his men must choose affects their calculations of self-interest, and cannot be described merely as setting boundaries within which these calculations must be applied. Rawls supposes, for example, that his men would inevitably choose conservative principles because this would be the only rational choice, in their ignorance, for self-interested men to make. But some actual men, aware of their own talents, might well prefer less conservative principles that would allow them to take advantage of the resources they know they have. Someone who considers the original position an argument for the conservative principles, therefore, is faced with this choice. If less conservative principles, like principles that favor named individuals, are to be ruled out as obviously unfair, then the argument for the conservative principles is complete at the outset, on grounds of obvious fairness alone. In that case neither the original position nor any considerations of self-interest it is meant to demonstrate play any role in the argument. But if less

conservative principles cannot be ruled out in advance as obviously unfair, then imposing ignorance on Rawls's men, so that they prefer the more conservative principles, cannot be explained simply as ruling out obviously unfair choices. And since this affects the antecedent self-interest of these men, the argument that the original position demonstrates the antecedent self-interest of actual men must therefore fail. This same dilemma can, of course, be constructed for each feature of the two principles.

I recognize that the argument thus far seems to ignore a distinctive feature of Rawls's methodology, which he describes as the technique of seeking a "reflective equilibrium" between our ordinary, unreflective moral beliefs and some theoretical structure that might unify and justify these ordinary beliefs.[1] It might now be said that the idea of an original position plays a part in this reflective equilibrium, which we will miss if we insist, as I have, on trying to find a more direct, one-way argument from the original position to the two principles of justice.

The technique of equilibrium does play an important role in Rawls's argument, and it is worth describing that technique briefly here. The technique assumes that Rawls's readers have a sense, which we draw upon in our daily life, that certain particular political arrangements or decisions like conventional trials, are just and others, like slavery, are unjust. It assumes, moreover, that we are each able to arrange these immediate intuitions or convictions in an order that designates some of them as more certain than others. Most people, for example, think that it is more plainly unjust for the state to execute innocent citizens of its own than to kill innocent foreign civilians in war. They might be prepared to abandon their position on foreign civilians in war, on the basis of some argument, but would be much more reluctant to abandon their view on executing innocent countrymen.

It is the task of moral philosophy, according to the technique of equilibrium, to provide a structure of principles that supports these immediate convictions about which we are more or less secure, with two goals in mind. First, this structure of principles must explain the convictions by showing the underlying assumptions they reflect; second, it must provide guidance in those cases about which we have either no convictions or weak or contradictory convictions. If we are unsure, for example, whether economic institutions that allow great disparity of wealth are unjust, we may turn to the principles that explain our confident convictions, and then apply these principles to that difficult issue.

---

[1] Pp. 48 ff.

But the process is not simply one of finding principles that accommodate our more-or-less settled judgments. These principles must support, and not merely account for, our judgments, and this means that the principles must have independent appeal to our moral sense. It might be, for example, that a cluster of familiar moral convictions could be shown to serve an undeserving policy—perhaps, that the standard judgments we make without reflection serve the purpose of maintaining one particular class in political power. But this discovery would not vouch for the principle of class egoism; on the contrary, it would discredit our ordinary judgments, unless some other principle of a more respectable sort could be found that also fits our intuitions, in which case it would be this principle and not the class-intent principle that our intuitions would recommend.

It might be that no coherent set of principles could be found that has independent appeal and that supports the full set of our immediate convictions; indeed it would be surprising if this were not often the case. If that does happen, we must compromise, giving way on both sides. We might relax, though we could not abandon, our initial sense of what might be an acceptable principle. We might come to accept, for example, after further reflection, some principle that seemed to us initially unattractive, perhaps the principle that men should sometimes be made to be free. We might accept this principle if we were satisfied that no less harsh principle could support the set of political convictions we were especially reluctant to abandon. On the other hand, we must also be ready to modify or adjust, or even to give up entirely, immediate convictions that cannot be accommodated by any principle that meets our relaxed standards; in adjusting these immediate convictions we will use our initial sense of which seem to us more and which less certain, though in principle no immediate conviction can be taken as immune from reinspection or abandonment if that should prove necessary. We can expect to proceed back and forth between our immediate judgments and the structure of explanatory principles in this way, tinkering first with one side and then the other, until we arrive at what Rawls calls the state of reflective equilibrium in which we are satisfied, or as much satisfied as we can reasonably expect.

It may well be that, at least for most of us, our ordinary political judgments stand in this relation of reflective equilibrium with Rawls's two principles of justice, or, at least, that they could be made to do so through the process of adjustment just described. It is nevertheless unclear how the idea of the original position fits into this structure or, indeed, why it has any role to play at all. The original position is not among the ordinary political convictions that we find we have, and

that we turn to reflective equilibrium to justify. If it has any role, it must be in the process of justification, because it takes its place in the body of theory we construct to bring our convictions into balance. But if the two principles of justice are themselves in reflective equilibrium with our convictions, it is unclear why we need the original position to supplement the two principles on the theoretical side of the balance. What can the idea contribute to a harmony already established?

We should consider the following answer. It is one of the conditions we impose on a theoretical principle, before we allow it to figure as a justification of our convictions, that the people the principle would govern would have accepted that principle, at least under certain conditions, if they had been asked, or at least that the principle can be shown to be in the antecedent interest of every such person. If this is so, then the original position plays an essential part in the process of justification through equilibrium. It is used to show that the two principles conform to this established standard of acceptability for political principles. At the same time, the fact that the two principles, which do conform to that standard, justify our ordinary convictions in reflective equilibrium reinforces our faith in the standard and encourages us to apply it to other issues of political or moral philosophy.

This answer does not advance the case that the original position furnishes an argument for the two principles, however; it merely restates the ideas we have already considered and rejected. It is certainly not part of our established political traditions or ordinary moral understanding that principles are acceptable only if they would be chosen by men in the particular predicament of the original position. It is, of course, part of these traditions that principles are fair if they have in fact been chosen by those whom they govern, or if they can at least be shown to be in their antecedent common interest. But we have already seen that the original position device cannot be used to support either of these arguments in favor of applying the two principles to contemporary politics. If the original position is to play any role in a structure of principles and convictions in reflective equilibrium, it must be by virtue of assumptions we have not yet identified.

It is time to reconsider an earlier assumption. So far I have been treating the original position construction as if it were either the foundation of Rawls's argument or an ingredient in a reflective equilibrium established between our political intuitions and his two principles of justice. But, in fact, Rawls does not treat the original position that way. He describes the construction in these words:

> I have emphasized that this original position is purely hypothet-

> ical. It is natural to ask why, if this agreement is never actually entered into, we should take any interest in these principles, moral or otherwise. The answer is that the conditions embodied in the description of the original position are ones that we do in fact accept. Or if we do not, then perhaps we can be persuaded to do so by philosophical reflection. Each aspect of the contractual situation can be given supporting grounds. . . . On the other hand, this conception is also an intuitive notion that suggests its own elaboration, so that led on by it we are drawn to define more clearly the standpoint from which we can best interpret moral relationships. We need a conception that enables us to envision our objective from afar: the intuitive notion of the original position is to do this for us.[2]

This description is taken from Rawls's first statement of the original position. It is recalled and repeated in the very last paragraph of the book.[3] It is plainly of capital importance, and it suggests that the original position, far from being the foundation of his argument, or an expository device for the technique of equilibrium, is one of the major substantive products of the theory as a whole. Its importance is reflected in another crucial passage. Rawls describes his moral theory as a type of psychology. He wants to characterize the structure of our (or, at least, one person's) capacity to make moral judgments of a certain sort, that is, judgments about justice. He thinks that the conditions embodied in the original position are the fundamental "principles governing our moral powers, or, more specifically, our sense of justice."[4] The original position is therefore a schematic representation of a particular mental process of at least some, and perhaps most, human beings, just as depth grammar, he suggests, is a schematic presentation of a different mental capacity.

All this suggests that the original position is an intermediate conclusion, a halfway point in a deeper theory that provides philosophical arguments for its conditions. In the next part of this essay I shall try to describe at least the main outlines of this deeper theory. I shall distinguish three features of the surface argument of the book—the technique of equilibrium, the social contract, and the original position itself—and try to discern which of various familiar philosophical principles or positions these represent.

First, however, I must say a further word about Rawls's exciting, if imprecise, idea that the principles of this deeper theory are constitutive

---

2 Pp. 21–22.

3 P. 587.

4 P. 51.

of our moral capacity. That idea can be understood on different levels of profundity. It may mean, at its least profound, that the principles that support the original position as a device for reasoning about justice are so widely shared and so little questioned within a particular community, for whom the book is meant, that the community could not abandon these principles without fundamentally changing its patterns of reasoning and arguing about political morality. It may mean, at its most profound, that these principles are innate categories of morality common to all men, imprinted in their neural structure, so that man could not deny these principles short of abandoning the power to reason about morality at all.

I shall be guided, in what follows, by the less profound interpretation, though what I shall say, I think, is consistent with the more profound. I shall assume, then, that there is a group of men and women who find, on reading Rawls, that the original position does strike them as a proper "intuitive notion" from which to think about problems of justice, and who would find it persuasive, if it could be demonstrated that the parties to the original position would in fact contract for the two principles he describes. I suppose, on the basis of experience and the literature, that this group contains a very large number of those who think about justice at all, and I find that I am a member myself. I want to discover the hidden assumptions that bend the inclinations of this group that way, and I shall do so by repeating the question with which I began. Why does Rawls's argument support his claim that his two principles are principles of justice? My answer is complex and it will take us, at times, far from his text, but not, I think, from its spirit.

II

A. Equilibrium

I shall start by considering the philosophical basis of the technique of equilibrium I just described. I must spend several pages in this way, but it is important to understand what substantive features of Rawls's deep theory are required by his method. This technique presupposes, as I said, a familiar fact about our moral lives. We all entertain beliefs about justice that we hold because they seem right, not because we have deduced or inferred them from other beliefs. We may believe in this way, for example, that slavery is unjust, and that the standard sort of trial is fair.

These different sorts of beliefs are, according to some philosophers, direct perceptions of some independent and objective moral facts. In the view of other philosophers they are simply subjective preferences, not unlike ordinary tastes, but dressed up in the language of justice

to indicate how important they seem to us. In any event, when we argue with ourselves or each other about justice we use these accustomed beliefs—which we call "intuitions" or "convictions"—in roughly the way Rawls's equilibrium technique suggests. We test general theories about justice against our own intuitions, and we try to confound those who disagree with us by showing how their own intuitions embarrass their own theories.

Suppose we try to justify this process by setting out a philosophical position about the connection between moral theory and moral intuition. The technique of equilibrium supposes what might be called a "coherence" theory of morality.[5] But we have a choice between two general models that define coherence and explain why it is required, and the choice between these is significant and consequential for our moral philosophy. I shall describe these two models, and then argue that the equilibrium technique makes sense on one but not the other.

I call the first a "natural" model. It presupposes a philosophical position that can be summarized in this way. Theories of justice, like Rawls's two principles, describe an objective moral reality; they are not, that is, created by men or societies but are rather discovered by them, as they discover laws of physics. The main instrument of this discovery is a moral faculty possessed by at least some men, which produces concrete intuitions of political morality in particular situations, like the intuition that slavery is wrong. These intuitions are clues to the nature and existence of more abstract and fundamental moral principles, as physical observations are clues to the existence and nature of fundamental physical laws. Moral reasoning or philosophy is a process of reconstructing the fundamental principles by assembling concrete judgments in the right order, as a natural historian reconstructs the shape of the whole animal from the fragments of its bones that he has found.

The second model is quite different. It treats intuitions of justice not as clues to the existence of independent principles, but rather as stipulated features of a general theory to be constructed, as if a sculptor set himself to carve the animal that best fit a pile of bones he happened to find together. This "constructive" model does not assume, as the natural model does, that principles of justice have some fixed, objective existence, so that descriptions of these principles must be true or false in some standard way. It does not assume that the animal it matches to the bones actually exists. It makes the different, and in some ways more complex, assumption that men and women have a responsibility to fit

---

[5] *See* Feinberg, *Justice, Fairness and Rationality*, 81 YALE L.J. 1004, 1018–21 (1972).

the particular judgments on which they act into a coherent program of action, or, at least, that officials who exercise power over other men have that sort of responsibility.

This second, constructive, model is not unfamiliar to lawyers. It is analogous to one model of common law adjudication. Suppose a judge is faced with a novel claim—for example, a claim for damages based on a legal right to privacy that courts have not heretofore recognized.[6] He must examine such precedents as seem in any way relevant to see whether any principles that are, as we might say, "instinct" in these precedents bear upon the claimed right to privacy. We might treat this judge as being in the position of a man arguing from moral intuitions to a general moral theory. The particular precedents are analogous to intuitions; the judge tries to reach an accommodation between these precedents and a set of principles that might justify them and also justify further decisions that go beyond them. He does not suppose, however, that the precedents are glimpses into a moral reality, and therefore clues to objective principles he ends by declaring. He does not believe that the principles are "instinct" in the precedents in that sense. Instead, in the spirit of the constructive model, he accepts these precedents as specifications for a principle that he must construct, out of a sense of responsibility for consistency with what has gone before.

I want to underline the important difference between the two models. Suppose that an official holds, with reasonable conviction, some intuition that cannot be reconciled with his other intuitions by any set of principles he can now fashion. He may think, for example, that it is unjust to punish an attempted murder as severely as a successful one, and yet be unable to reconcile that position with his sense that a man's guilt is properly assessed by considering only what he intended, and not what actually happened. Or he may think that a particular minority race, as such, is entitled to special protection, and be unable to reconcile that view with his view that distinctions based on race are inherently unfair to individuals. When an official is in this position the two models give him different advice.

The natural model supports a policy of following the troublesome intuition, and submerging the apparent contradiction, in the faith that a more sophisticated set of principles, which reconciles that intuition, does in fact exist though it has not yet been discovered. The official,

---

[6] I have here in mind the fairness argument of Brandeis and Warren. *See* Brandeis & Warren, *The Right to Privacy*, 4 HARV. L. REV. 193 (1890), which is a paradigm of argument in the constructive model.

according to this model, is in the position of the astronomer who has clear observational data that he is as yet unable to reconcile in any coherent account, for example, of the origin of the solar system. He continues to accept and employ his observational data, placing his faith in the idea that some reconciling explanation does exist though it has not been, and for all he knows may never be, discovered by men.

The natural model supports this policy because it is based on a philosophical position that encourages the analogy between moral intuitions and observational data. It makes perfect sense, on that assumption, to suppose that direct observations, made through a moral faculty, have outstripped the explanatory powers of those who observe. It also makes sense to suppose that some correct explanation, in the shape of principles of morality, does in fact exist in spite of this failure; if the direct observations are sound, some explanation must exist for why matters are as they have been observed to be in the moral universe, just as some explanation must exist for why matters are as they have been observed to be in the physical universe.

The constructive model, however, does not support the policy of submerging apparent inconsistency in the faith that reconciling principles must exist. On the contrary, it demands that decisions taken in the name of justice must never outstrip an official's ability to account for these decisions in a theory of justice, even when such a theory must compromise some of his intuitions. It demands that we act on principle rather than on faith. Its engine is a doctrine of responsibility that requires men to integrate their intuitions and subordinate some of these, when necessary, to that responsibility. It presupposes that articulated consistency, decisions in accordance with a program that can be made public and followed until changed, is essential to any conception of justice. An official in the position I describe, guided by this model, must give up his apparently inconsistent position; he must do so even if he hopes one day, by further reflection, to devise better principles that will allow all his initial convictions to stand as principles.[7]

The constructive model does not presuppose scepticism or relativism. On the contrary, it assumes that the men and women who reason within the model will each hold sincerely the convictions they bring to it, and that this sincerity will extend to criticizing as unjust political acts or systems that offend the most profound of these. The model does

---

7 The fairness debate between Professor Wechsler, *Toward Neutral Principles in Constitutional Law*, 73 HARV. L. REV. 1 (1959), and his critics may be illuminated by this distinction. Wechsler proposes a constructive model for constitutional adjudication, while those who favor a more tentative or intuitive approach to constitutional law are following the material model.

not deny, any more than it affirms, the objective standing of any of these convictions; it is therefore consistent with, though as a model of reasoning it does not require, the moral ontology that the natural model presupposes.

It does not require that ontology because its requirements are independent of it. The natural model insists on consistency with conviction, on the assumption that moral intuitions are accurate observations; the requirement of consistency follows from that assumption. The constructive model insists on consistency with conviction as an independent requirement, flowing not from the assumption that these convictions are accurate reports, but from the different assumption that it is unfair for officials to act except on the basis of a general public theory that will constrain them to consistency, provide a public standard for testing or debating or predicting what they do, and not allow appeals to unique intuitions that might mask prejudice or self-interest in particular cases. The constructive model requires coherence, then, for independent reasons of political morality; it takes convictions held with the requisite sincerity as given, and seeks to impose conditions on the acts that these intuitions might be said to warrant. If the constructive model is to constitute morality, in either of the senses I have distinguished, these independent reasons of political morality are at the heart of our political theories.

The two models, therefore, represent different standpoints from which theories of justice might be developed. The natural model, we might say, looks at intuitions from the personal standpoint of the individual who holds them, and who takes them to be discrete observations of moral reality. The constructive model looks at these intuitions from a more public standpoint; it is a model that someone might propose for the governance of a community each of whose members has strong convictions that differ, though not too greatly, from the convictions of others.

The constructive model is appealing, from this public standpoint, for an additional reason. It is well suited to group consideration of problems of justice, that is, to developing a theory that can be said to be the theory of a community rather than of particular individuals, and this is an enterprise that is important, for example, in adjudication. The range of initial convictions to be assessed can be expanded or contracted to accommodate the intuitions of a larger or smaller group, either by including all convictions held by any members, or by excluding those not held by all, as the particular calculation might warrant. This process would be self-destructive on the natural model,

because every individual would believe that either false observations were being taken into account or accurate observations disregarded, and hence that the inference to objective morality was invalid. But on the constructive model that objection would be unavailable; the model, so applied, would be appropriate to identify the program of justice that best accommodates the community's common convictions, for example, with no claim to a description of an objective moral universe.

Which of these two models, then, better supports the technique of equilibrium? Some commentators seem to have assumed that the technique commits Rawls to the natural model.[8] But the alliance between that model and the equilibrium technique turns out to be only superficial; when we probe deeper we find that they are incompatible. In the first place, the natural model cannot explain one distinctive feature of the technique. It explains why our theory of justice must fit our intuitions about justice, but it does not explain why we are justified in amending these intuitions to make the fit more secure.

Rawls's notion of equilibrium, as I said earlier, is a two-way process; we move back and forth between adjustments to theory and adjustments to conviction until the best fit possible is achieved. If my settled convictions can otherwise be captured by, for example, a straightforward utilitarian theory of justice, that may be a reason, within the technique, for discarding my intuition that slavery would be wrong even if it advanced utility. But on the natural model this would be nothing short of cooking the evidence, as if a naturalist rubbed out the footprints that embarrassed his efforts to describe the animal that left them, or the astronomer just set aside the observations that his theory could not accommodate.

We must be careful not to lose this point in false sophistication about science. It is common to say—Rawls himself draws the comparison[9]—that scientists also adjust their evidence to achieve a smooth set of explanatory principles. But if this is true at all, their procedures are very different from those recommended by the technique of equilibrium. Consider, to take a familiar example, optical illusions or hallucinations. It is perfectly true that the scientist who sees water in the sand does not say that the pond was really there until he arrived at it, so that physics must be revised to provide for disappearing water; on the contrary, he uses the apparent disappearing as evidence of an illu-

8 *See e.g.*, Hare, *Rawls' Theory of Justice—I*, 23 PHILOSOPHICAL QUARTERLY 144 (1973).

9 Rawls draws attention to the distinction. P. 49.

sion, that is, as evidence that, contrary to his observation, there was never any water there at all.

The scientists, of course, cannot leave the matter at that. He cannot dismiss mirages unless he supplements the laws of physics with laws of optics that explain them. It may be that he has, in some sense, a choice amongst competing sets of explanations of all his observations taken together. He may have a choice, for example, between either treating mirages as physical objects of a special sort and then amending the laws of physics to allow for disappearing objects of this sort, or treating mirages as optical illusions and then developing laws of optics to explain such illusions. He has a choice in the sense that his experience does not absolutely force either of these explanations upon him; the former is a possible choice, though it would require wholesale revision of both physics and common sense to carry it off.

This is, I take it, what is meant by philosophers like Quine who suppose that our concepts and our theories face our experience as a whole, so that we might react to recalcitrant or surprising experience by making different revisions at different places in our theoretical structures if we wish.[10] Regardless of whether this is an accurate picture of scientific reasoning, it is not a picture of the procedure of equilibrium, because this procedure argues not simply that alternative structures of principle are available to explain the same phenomena, but that some of the phenomena, in the form of moral convictions, may simply be ignored the better to serve some particular theory.

It is true that Rawls sometimes describes the procedure in a more innocent way. He suggests that if our tentative theories of justice do not fit some particular intuition, this should act as a warning light requiring us to reflect on whether the conviction is really one we hold.[11] If my convictions otherwise support a principle of utility, but I feel that slavery would be unjust even if utility were improved, I might think about slavery again, in a calmer way, and this time my intuitions might be different and consistent with that principle. In this case, the initial inconsistency is used as an occasion for reconsidering the intuition, but not as a reason for abandoning it.

Still, this need not happen. I might continue to receive the former intuition, no matter how firmly I steeled myself against it. In that case the procedure nevertheless authorizes me to set it aside if that is required to achieve the harmony of equilibrium. But if I do, I am not offering an alternative account of the evidence, but simply disregarding

10 W. V. Quine, *Two Dogmas of Empiricism*, in FROM A LOGICAL POINT OF VIEW 20 (2d ed. rev. 1964).

11 P. 48.

it. Someone else, whose intuitions are different, may say that mine are distorted, perhaps because of some childhood experience, or because I am insufficiently imaginative to think of hypothetical cases in which slavery might actually improve utility. He may say, that is, that my sensibilities are defective here, so that my intuitions are not genuine perceptions of moral reality, and may be set aside like the flawed reports of a color-blind man.

But I cannot accept that about myself, as an explanation for my own troublesome convictions, so long as I hold these convictions and they seem to me sound, indistinguishable in their moral quality from my other convictions. I am in a different position from the color-blind man who need only come to understand that others' perceptions differ from his. If I believe that my intuitions are a direct report from some moral reality, I cannot accept that one particular intuition is false until I come to feel or sense that it is false. The bare fact that others disagree, if they do, may be an occasion for consulting my intuitions again, but if my convictions remain the same, the fact that others may explain them in a different way cannot be a reason for my abandoning them, instead of retaining them in the faith that a reconciliation of these with my other convictions does in fact exist.

Thus, the natural model does not offer a satisfactory explanation of the two-way feature of equilibrium. Even if it did, however, it would leave other features of that technique unexplained; it would leave unexplained, for example, the fact that the results of the technique, at least in Rawls's hands, are necessarily and profoundly practical. Rawlsian men and women in the original position seek to find principles that they and their successors will find it easy to understand and publicize and observe; principles otherwise appealing are to be rejected or adjusted because they are too complex or are otherwise impractical in this sense. But principles of justice selected in this spirit are compromises with infirmity, and are contingent in the sense that they will change as the general condition and education of people change. This seems inconsistent with the spirit, at least, of the natural model, according to which principles of justice are timeless features of some independent moral reality, to which imperfect men and women must attempt to conform at best they can.

The equilibrium technique, moreover, is designed to produce principles that are relative in at least two ways. First, it is designed to select the best theory of justice from a list of alternative theories that must not only be finite, but short enough to make comparisons among them feasible. This limitation is an important one; it leads Rawls himself to

say that he has no doubt that an initial list of possible theories expanded well beyond the list he considers would contain a better theory of justice than his own two principles.[12] Second, it yields results that are relative to the area of initial agreement among those who jointly conduct the speculative experiments it recommends. It is designed, as Rawls says, to reconcile men who disagree by fixing on what is common ground among them.[13] The test concededly will yield different results for different groups, and for the same group at different times, as the common ground of confident intuition shifts.

If the equilibrium technique were used within the natural model, the authority of its conclusions would be seriously compromised by both forms of relativism. If the equilibrium argument for Rawls's two principles, for example, shows only that a better case can be made for them than for any other principles on a restricted short list, and if Rawls himself is confident that further study would produce a better theory, then we have very little reason to suppose that these two principles are an accurate description of moral reality. It is hard to see, on the natural model, why they then should have any authority at all.

Indeed, the argument provides no very good ground for supposing even that the two principles are a better description of moral reality than other theories on the short list. Suppose we are asked to choose, among five theories of justice, the theory that best unites our convictions in reflective equilibrium, and we pick, from among these, the fifth. Let us assume that there is some sixth theory that we would have chosen had it appeared on the list. This sixth theory might be closer to, for example, the first on our original list than to the fifth, at least in the following sense: over a long term, a society following the first might reach more of the decisions that a society following the sixth would reach than would a society following the fifth.

Suppose, for example, that our original list included, as available theories of justice, classical utilitarianism and Rawls's two principles, but did not include average utilitarianism. We might have rejected classical utilitarianism on the ground that the production of pleasure for its own sake, unrelated to any increase in the welfare of particular human beings or other animals, makes little sense, and then chosen Rawls's two principles as the best of the theories left. We might nevertheless have chosen average utilitarianism as superior to the two principles, if it had been on the list, because average utilitarianism does not suppose that just any increase in the total quantity of pleasure is good.

---

12 P. 581.

13 Pp. 580–81.

But classical utilitarianism, which we rejected, might be closer to average utilitarianism, which we would have chosen if we could have, than are the two principles which we did choose. It might be closer, in the sense described, because it would dictate more of the particular decisions that average utilitarianism would require, and thus be a better description of ultimate moral reality, than would the two principles. Of course, average utilitarianism might itself be rejected in a still larger list, and the choice we should then make might indicate that another member of the original list was better than either classical utilitarianism or the two principles.

The second sort of relativism would be equally damaging on the natural model, for reasons I have already explained. If the technique of equilibrium is used by a single person, and the intuitions allowed to count are just his and all of his, then the results may be authoritative for him. Others, whose intuitions differ, will not be able to accept his conclusions, at least in full, but he may do so himself. If, however, the technique is used in a more public way, for example, by fixing on what is common amongst the intuitions of a group, then the results will be those that no one can accept as authoritative, just as no one could accept as authoritative a scientific result reached by disregarding what he believed to be evidence at least as pertinent as the evidence used.

So the natural model turns out to be poor support for the equilibrium technique. None of the difficulties just mentioned count, however, if we assume the technique to be in the service of the constructive model. It is, within that model, a reason for rejecting even a powerful conviction that it cannot be reconciled with other convictions by a plausible and coherent set of principles; the conviction is rejected not as a false report, but simply as ineligible within a program that meets the demands of the model. Nor does either respect in which the technique is relative embarrass the constructive model. It is not an embarrassment that some theory not considered might have been deemed superior if it had been considered. The model requires officials or citizens to proceed on the best program they can now fashion, for reasons of consistency that do not presuppose, as the natural model does, that the theory chosen is in any final sense true. It does not undermine a particular theory that a different group, or a different society, with different culture and experience, would produce a different one. It may call into question whether any group is entitled to treat its moral intuitions as in any sense objective or transcendental, but not that a particular society, which does treat particular convictions in that way, is therefore required to follow them in a principled way.

I shall assume, therefore, at least tentatively, that Rawls's methodol-

ogy presupposes the constructive model of reasoning from particular convictions to general theories of justice, and I shall use that assumption in my attempt to show the further postulates of moral theory that lie behind his theory of justice.

B. The Contract

I come, then, to the second of the three features of Rawls's methodology that I want to discuss, which is the use he makes of the old idea of a social contract. I distinguish, as does Rawls, the general idea that an imaginary contract is an appropriate device for reasoning about justice, from the more specific features of the original position, which count as a particular application of that general idea. Rawls thinks that all theories that can be seen to rest on a hypothetical social contract of some sort are related and are distinguished as a class from theories that cannot; he supposes, for example, that average utilitarianism, which can be seen as the product of a social contract on a particular interpretation, is more closely related to his own theory than either is to classical utilitarianism, which cannot be seen as the product of a contract on any interpretation.[14] In the next section I shall consider the theoretical basis of the original position. In this section I want to consider the basis of the more general idea of the contract itself.

Rawls says that the contract is a powerful argument for his principles because it embodies philosophical principles that we accept, or would accept if we thought about them. We want to find out what these principles are, and we may put our problem this way. The two principles comprise a theory of justice that is built up from the hypothesis of a contract. But the contract cannot sensibly be taken as the fundamental premise or postulate of that theory, for the reasons I described in the first part of this article. It must be seen as a kind of halfway point in a larger argument, as itself the product of a deeper political theory that argues for the two principles *through* rather than *from* the contract. We must therefore try to identify the features of a deeper theory that would recommend the device of a contract as the engine for a theory of justice, rather than the other theoretical devices Rawls mentions, like the device of the impartial spectator.[15]

We shall find the answer, I think, if we attend to and refine the familiar distinction philosophers make between two types of moral theories, which they call teleological theories and deontological theories.[16] I shall argue that any deeper theory that would justify Rawls's

14 Chapter 30.

15 Pp. 144 ff.

16 Rawls defines these terms at pp. 24–25 and 30.

use of the contract must be a particular form of deontological theory, a theory that takes the idea of rights so seriously as to make them fundamental in political morality. I shall try to show how such a theory would be distinguished, as a type, from other types of political theories, and why only such a theory could give the contract the role and prominence Rawls does.

I must begin this argument, however, by explaining how I shall use some familiar terms. (1) I shall say that some state of affairs is a *goal* within a particular political theory if it counts in favor of a political act, within that theory, that the act will advance or preserve that state of affairs, and counts against an act that it will retard or threaten it. Goals may be relatively specific, like full employment or respect for authority, or relatively abstract, like improving the general welfare, advancing the power of a particular nation, or creating a utopian society according to a particular concept of human goodness or of the good life. (2) I shall say that an individual has a *right* to a particular political act, within a political theory, if the failure to provide that act, when he calls for it, would be unjustified within that theory even if the goals of the theory would, on the balance, be disserviced by that act. The strength of a particular right, within a particular theory, is a function of the degree of disservice to the goals of the theory, beyond a mere disservice on the whole, that is necessary to justify refusing an act called for under the right. In the popular political theory apparently prevailing in the United States, for example, individuals have rights to free public speech on political matters and to a certain minimum standard of living, but neither right is absolute and the former is much stronger than the latter. (3) I shall say that an individual has a *duty* to act in a particular way, within a political theory, if a political decision constraining such act is justified within that theory notwithstanding that no goal of the system would be served by that decision. A theory may provide, for example, that individuals have a duty to worship God, even though it does not stipulate any goal served by requiring them to do so.[17]

The three concepts I have described work in different ways, but they all serve to justify or to condemn, at least pro tanto, particular political decisions. In each case, the justification provided by citing a goal, a right, or a duty is in principle complete, in the sense that nothing need be added to make the justification effective, if it is not undermined by some competing considerations. But, though such a justification is in

[17] I do not count, as goals, the goal of respecting rights or enforcing duties. In this and other apparent ways my use of the terms I define is narrower than ordinary language permits.

this sense complete, it need not, within the theory, be ultimate. It remains open to ask why the particular goal, right, or duty is itself justified, and the theory may provide an answer by deploying a *more basic* goal, right, or duty that is served by accepting this less basic goal, right, or duty as a complete justification in particular cases.

A particular goal, for example, might be justified as contributing to a more basic goal; thus, full employment might be justified as contributing to greater average welfare. Or a goal might be justified as serving a more basic right or duty; a theory might argue, for example, that improving the gross national product, which is a goal, is necessary to enable the state to respect the rights of individuals to a decent minimum standard of living, or that improving the efficiency of the police process is necessary to enforce various individual duties not to sin. On the other hand, rights and duties may be justified on the ground that, by acting as a complete justification on particular occasions, they, in fact serve more fundamental goals; the duty of individuals to drive carefully may be justified, for example, as serving the more basic goal of improving the general welfare. This form of justification does not, of course, suggest that the less basic right or duty itself justifies political decisions only when these decisions, considered one by one, advance the more basic goal. The point is rather the familiar one of rule utilitarianism, that treating the right or duty as a complete justification in particular cases, without reference to the more basic goal, will in fact advance the goal in the long run.

So goals can be justified by other goals or by rights or duties, and rights or duties can be justified by goals. Rights and duties can also be justified, of course, by other, more fundamental duties or rights. Your duty to respect my privacy, for example, may be justified by my right to privacy. I do not mean merely that rights and duties may be correlated, as opposite sides of the same coin. That may be so when, for example, a right and the corresponding duty are justified as serving a more fundamental goal, as when your right to property and my corresponding duty not to trespass are together justified by the more fundamental goal of socially efficient land use. In many cases, however, corresponding rights and duties are not correlative, but one is derivative from the other, and it makes a difference which is derivative from which. There is a difference between the idea that you have a duty not to lie to me because I have a right not to be lied to, and the idea that I have a right that you not lie to me because you have a duty not to tell lies. In the first case I justify a duty by calling attention to a right; if I intend any further justification it is the right that I must justify, and I cannot do so by calling attention to the duty. In the second case it is the other

way around. The difference is important because, as I shall shortly try to show, a theory that takes rights as fundamental is a theory of a different character from one that takes duties as fundamental.

Political theories will differ from one another, therefore, not simply in the particular goals, rights, and duties each sets out, but also in the way each connects the goals, rights, and duties it employs. In a well-formed theory some consistent set of these, internally ranked or weighted, will be taken as fundamental or ultimate within the theory. It seems reasonable to suppose that any particular theory will give ultimate pride of place to just one of these concepts; it will take some overriding goal, or some set of fundamental rights, or some set of transcendent duties, as fundamental, and show other goals, rights, and duties as subordinate and derivative.[18]

We may therefore make a tentative initial classification of the political theories we might produce, on the constructive model, as deep theories that might contain a contract as an intermediate device. Such a theory might be *goal-based,* in which case it would take some goal, like improving the general welfare, as fundamental; it might be *right-based,* taking some right, like the right of all men to the greatest possible overall liberty, as fundamental; or it might be *duty-based,* taking some duty, like the duty to obey God's will as set forth in the Ten Commandments, as fundamental. It is easy to find examples of pure, or nearly pure, cases of each of these types of theory. Utilitarianism is, as my example suggested, a goal-based theory; Kant's categorical imperatives compose a duty-based theory; and Tom Paine's theory of revolution is right-based.

Theories within each of these types are likely to share certain very general characteristics. The types may be contrasted, for example, by comparing the attitudes they display towards individual choice and conduct. Goal-based theories are concerned with the welfare of any particular individual only in so far as this contributes to some state of affairs stipulated as good quite apart from his choice of that state of affairs. This is plainly true of totalitarian goal-based theories, like fascism, that take the interest of a political organization as fundamental. It is also true of the various forms of utilitarianism, because, though they count up the impact of political decisions on distinct individuals, and are in this way concerned with individual welfare, they merge these impacts into overall totals or averages and take the improvement of these totals or averages as desirable quite apart from the decision of any individual that it is. It is also true of perfectionist theories, like Aristotle's, that impose upon individuals an ideal of excellence and take the goal of politics to be the culture of such excellence.

18 But an "intuitionist" theory, as Rawls uses that term, need not. *See* p. 34.

Right-based and duty-based theories, on the other hand, place the individual at the center, and take his decision or conduct as of fundamental importance. But the two types put the individual in a different light. Duty-based theories are concerned with the moral quality of his acts, because they suppose that it is wrong, without more, for an individual to fail to meet certain standards of behavior. Kant thought that it was wrong to tell a lie no matter how beneficial the consequences, not because having this practice promoted some goal, but just because it was wrong. Right-based theories are, in contrast, concerned with the independence rather than the conformity of individual action. They presuppose and protect the value of individual thought and choice. Both types of theory make use of the idea of moral rules, codes of conduct to be followed, on individual occasions, without consulting self-interest. Duty-based theories treat such codes of conduct as of the essence, whether set by society to the individual or by the individual to himself. The man at their center is the man who must conform to such a code, or be punished or corrupted if he does not. Right-based theories, however, treat codes of conduct as instrumental, perhaps necessary to protect the rights of others, but having no essential value in themselves. The man at their center is the man who benefits from others' compliance, not the man who leads the life of virtue by complying himself.

We should, therefore, expect that the different types of theories would be associated with different metaphysical or political temperaments, and that one or another would be dominant in certain sorts of political economy. Goal-based theories, for example, seem especially compatible with homogeneous societies, or those at least temporarily united by an urgent, overriding goal, like self-defense or economic expansion. We should also expect that these differences between types of theory would find echoes in the legal systems of the communities they dominate. We should expect, for example, that a lawyer would approach the question of punishing moral offenses through the criminal law in a different way if his inchoate theory of justice were goal-, right- or duty-based. If his theory were goal-based he would consider the full effect of enforcing morality upon his overriding goal. If this goal were utilitarian, for example, he would entertain, though he might, in the end, reject, Lord Devlin's arguments that the secondary effects of punishing immorality may be beneficial.[19] If his theory were duty-based, on the other hand, he would see the point of the argument, commonly called retributive, that since immorality is wrong the state must punish it even if it harms no one. If his theory were right-based, however, he

[19] *See* Dworkin, *Lord Devlin and the Enforcement of Morals*, 75 YALE L.J. 986 (1966).

would reject the retributive argument, and judge the utilitarian argument against the background of his own assumption that individual rights must be served even at some cost to the general welfare.

All this is, of course, superficial and trivial as ideological sociology. My point is only to suggest that these differences in the character of a political theory are important quite apart from the details of position that might distinguish one theory from another of the same character. It is for this reason that the social contract is so important a feature of Rawls's methodology. It signals that his deep theory is a right-based theory, rather than a theory of either of the other two types.

The social contract provides every potential party with a veto: unless he agrees, no contract is formed. The importance, and even the existence, of this veto is obscured in the particular interpretation of the contract that constitutes the original position. Since no one knows anything about himself that would distinguish him from anyone else, he cannot rationally pursue any interest that is different. In these circumstances nothing turns on each man having a veto, or, indeed, on there being more than one potential party to the contract in the first place. But the original position is only one interpretation of the contract, and in any other interpretation in which the parties do have some knowledge with which to distinguish their situation or ambitions from those of others, the veto that the contract gives each party becomes crucial. The force of the veto each individual has depends, of course, upon his knowledge, that is to say, the particular interpretation of the contract we in the end choose. But the fact that individuals should have any veto at all is in itself remarkable.

It can have no place in a purely goal-based theory, for example. I do not mean that the parties to a social contract could not settle on a particular social goal and make that goal henceforth the test of the justice of political decisions. I mean that no goal-based theory could make a contract the proper device for deciding upon a principle of justice in the first place; that is, the deep theory we are trying to find could not itself be goal-based.

The reason is straightforward. Suppose some particular overriding goal, like the goal of improving the average welfare in a community, or increasing the power and authority of a state, or creating a utopia according to a particular conception of the good, is taken as fundamental within a political theory. If any such goal is fundamental, then it authorizes such distribution of resources, rights, benefits, and burdens within the community as will best advance that goal, and condemns any other. The contract device, however, which supposes each individual to pursue his own interest and gives each a veto on the collective de-

cision, applies a very different test to determine the optimum distribution. It is designed to produce the distribution that each individual deems in his own best interest, given his knowledge under whatever interpretation of the contract is specified, or at least to come as close to that distribution as he thinks he is likely to get. The contract, therefore, offers a very different test of optimum distribution than a direct application of the fundamental goal would dictate. There is no reason to suppose that a system of individual vetoes will produce a good solution to a problem in which the fairness of a distribution, considered apart from the contribution of the distribution to an overall goal, is meant to count for nothing.

It might be, of course, that a contract would produce the result that some fundamental goal dictates. Some critics, in fact, think that men in the original position, Rawls's most favored interpretation of the contract, would choose a theory of justice based on principles of average utility, that is, just the principles that a deep theory stipulating the fundamental goal of average utility would produce.[20] But if this is so, it is either because of coincidence or because the interpretation of the contract has been chosen to produce this result; in either case the contract is supererogatory, because the final result is determined by the fundamental goal and the contract device adds nothing.

One counterargument is available. Suppose it appears that the fundamental goal will in fact be served only if the state is governed in accordance with principles that all men will see to be, in some sense, in their own interest. If the fundamental goal is the aggrandizement of the state, for example, it may be that this goal can be reached only if the population does not see that the government acts for this goal, but instead supposes that it acts according to principles shown to be in their individual interests through a contract device; only if they believe this will they work in the state's interest at all. We cannot ignore this devious, if unlikely, argument, but it does not support the use that Rawls makes of the contract. The argument depends upon a deception, like Sidgewick's famous argument that utilitarianism can best be served by keeping the public ignorant of that theory.[21] A theory that includes such a deception is ineligible on the constructivist model we are pursuing, because our aim, on that model, is to develop a theory that unites our convictions and can serve as a program for public action; publicity is as much a requirement of our deep theory as of the conception of justice that Rawls develops within it.

---

20 John Mackie presented a forceful form of this argument to an Oxford seminar in the fall of 1972.

21 H. SIDGEWICK, THE METHODS OF ETHICS 489 ff. (7th ed. 1907).

So a goal-based deep theory cannot support the contract, except as a useless and confusing appendage. Neither can a duty-based deep theory, for much the same reasons. A theory that takes some duty or duties to be fundamental offers no ground to suppose that just institutions are those seen to be in everyone's self-interest under some description. I do not deny, again, that the parties to the contract may decide to impose certain duties upon themselves and their successors, just as they may decide to adopt certain goals, in the exercise of their judgment of their own self-interest. Rawls describes the duties they would impose upon themselves under his most favored interpretation, the original position, and calls these natural duties.[22] But this is very different from supposing that the deep theory, which makes this decision decisive of what these duties are, can itself be duty-based.

It is possible to argue, of course, as many philosophers have, that a man's self-interest lies in doing his duty under the moral law, either because God will punish him otherwise, or because fulfilling his role in the natural order is his most satisfying activity, or, as Kant thought, because only in following rules he could consistently wish universal can he be free. But that says a man's duties define his self-interest, and not the other way round. It is an argument not for deciding upon a man's particular duties by letting him consult his own interest, but rather for his setting aside any calculations of self-interest except calculations of duty. It could not, therefore, support the role of a Rawlsian contract in a duty-based deep theory.

It is true that if a contract were a feature of a duty-based deep theory, an interpretation of the contract could be chosen that would dissolve the apparent conflict between self-interest and duty. It might be a feature of the contract situation, for example, that all parties accepted the idea just mentioned, that their self-interest lay in ascertaining and doing their duty. This contract would produce principles that accurately described their duties, at least if we add the supposition that they are proficient, for some reason, in discovering what their duties are. But then, once again, we have made the contract supererogatory, a march up the hill and then back down again. We would have done better simply to work out principles of justice from the duties the deep theory takes as fundamental.

The contract does, however, make sense in a right-based deep theory. Indeed, it seems a natural development of such a theory. The basic idea of a right-based theory is that distinct individuals have interests that they are entitled to protect if they so wish. It seems natural, in develop-

22 Chapter 19.

ing such a theory, to try to identify the institutions an individual would veto in the exercise of whatever rights are taken as fundamental. The contract is an excellent device for this purpose, for at least two reasons. First, it allows us to distinguish between a veto in the exercise of these rights and a veto for the sake of some interest that is not so protected, a distinction we can make by adopting an interpretation of the contract that reflects our sense of what these rights are. Second, it enforces the requirements of the constructive model of argument. The parties to the contract face a practical problem; they must devise a constitution from the options available to them, rather than postponing their decision to a day of later moral insight, and they must devise a program that is both practical and public in the sense I have described.

It seems fair to assume, then, that the deep theory behind the original position must be a right-based theory of some sort. There is another way to put the point, which I have avoided until now. It must be a theory that is based on the concept of rights that are *natural*, in the sense that they are not the product of any legislation, or convention, or hypothetical contract. I have avoided that phrase because it has, for many people, disqualifying metaphysical associations. They think that natural rights are supposed to be spectral attributes worn by primitive men like amulets, which they carry into civilization to ward off tyranny. Mr. Justice Black, for example, thought it was a sufficient refutation of a judicial philosophy he disliked simply to point out that it seemed to rely on this preposterous notion.[23]

But on the constructive model, at least, the assumption of natural rights is not a metaphysically ambitious one. It requires no more than the hypothesis that the best political program, within the sense of that model, is one that takes the protection of certain individual choices as fundamental, and not properly subordinated to any goal or duty or combination of these. This requires no ontology more dubious or controversial than any contrary choice of fundamental concepts would be and, in particular, no more than the hypothesis of a fundamental goal that underlies the various popular utilitarian theories would require. Nor is it disturbing that a Rawlsian deep theory makes these rights natural rather than legal or conventional. Plainly, any right-based theory must presume rights that are not simply the product of deliberate legislation or explicit social custom, but are independent grounds for judging legislation and custom. On the constructive model, the assumption that rights are in this sense natural is simply one assumption to be made and examined for its power to unite and explain our political

23 Griswold v. Connecticut, 381 U.S. 479, 507 (1964) (dissenting opinion).

convictions, one basic programatic decision to submit to this test of coherence and experience.

C. The Original Position

I said that the use of a social contract, in the way that Rawls uses it, presupposes a deep theory that assumes natural rights. I want now to describe, in somewhat more detail, how the device of a contract applies that assumption. It capitalizes on the idea, mentioned earlier, that some political arrangements might be said to be in the antecedent interest of every individual even though they are not, in the event, in his actual interest.

Everyone whose consent is necessary to a contract has a veto over the terms of that contract, but the worth of that veto, to him, is limited by the fact that his judgment must be one of antecedent rather than actual self-interest. He must commit himself, and so abandon his veto, at a time when his knowledge is sufficient only to allow him to estimate the best odds, not to be certain of his bet. So the contract situation is in one way structurally like the situation in which an individual with specific political rights confronts political decisions that may disadvantage him. He has a limited, political right to veto these, a veto limited by the scope of the rights he has. The contract can be used as a model for the political situation by shaping the degree or character of a party's ignorance in the contractual situation so that this ignorance has the same force on his decision as the limited nature of his rights would have in the political situation.

This shaping of ignorance to suit the limited character of political rights is most efficiently done simply by narrowing the individual goals that the parties to the contract know they wish to pursue. If we take Hobbes's deep theory, for example, to propose that men have a fundamental natural right to life, so that it is wrong to take their lives, even for social goals otherwise proper, we should expect a contract situation of the sort he describes. Hobbes's men and women, in Rawls's phrase, have lexically ordered security of life over all other individual goals; the same situation would result if they were simply ignorant of any other goals they might have and unable to speculate about the chances that they have any particular one or set of these.

The ignorance of the parties in the original position might thus be seen as a kind of limiting case of the ignorance that can be found, in the form of a distorted or eccentric ranking of interests, in classical contract theories and that is natural to the contract device. The original position is a limiting case because Rawls's men are not simply ignorant of interests beyond a chosen few; they are ignorant of all the interests they

have. It would be wrong to suppose that this makes them incapable of any judgments of self-interest. But the judgments they make must nevertheless be very abstract; they must allow for any combination of interests, without the benefit of any supposition that some of these are more likely than others.

The basic right of Rawls's deep theory, therefore, cannot be a right to any particular individual goal, like a right to security of life, or a right to lead a life according to a particular conception of the good. Such rights to individual goals may be produced by the deep theory, as rights that men in the original position would stipulate as being in their best interest. But the original position cannot itself be justified on the assumption of such a right, because the parties to the contract do not know that they have any such interest or rank it lexically ahead of others.

So the basic right of Rawls's deep theory must be an abstract right, that is, not a right to any particular individual goal. There are two candidates, within the familiar concepts of political theory, for this role. The first is the right to liberty, and it may strike many readers as both plausible and comforting to assume that Rawls's entire structure is based on the assumption of a fundamental natural right to liberty—plausible because the two principles that compose his theory of justice give liberty an important and dominant place, and comforting because the argument attempting to justify that place seems uncharacteristically incomplete.[24]

Nevertheless, the right to liberty cannot be taken as the fundamental right in Rawls's deep theory. Suppose we define general liberty as the overall minimum possible constraints, imposed by government or by other men, on what a man might want to do.[25] We must then distinguish this general liberty from particular liberties, that is, freedom from such constraints on particular acts thought specially important, like participation in politics. The parties to the original position certainly have, and know that they have, an interest in general liberty, because general liberty will, pro tanto, improve their power to achieve any particular goals they later discover themselves to have. But the qualification is important, because they have no way of knowing that general liberty will in fact improve this power overall, and every reason to suspect that it will not. They know that they might have other interests, beyond general liberty, that can be protected only by political constraints on acts of others.

---

24 *See* Hart, *Rawls on Liberty and Its Priority*, 40 U. CHI. L. REV. 534 (1973).

25 *Cf.* Rawls's definition of liberty at p. 202.

So if Rawlsian men must be supposed to have a right to liberty of some sort, which the contract situation is shaped to embody, it must be a right to particular liberties. Rawls does name a list of basic liberties, and it is these that his men do choose to protect through their lexically ordered first principle of justice.[26] But Rawls plainly casts this principle as the product of the contract rather than as a condition of it. He argues that the parties to the original position would select these basic liberties to protect the basic goods they decide to value, like self-respect, rather than taking these liberties as goals in themselves. Of course they might, in fact, value the activities protected as basic liberties for their own sake, rather than as means to some other goal or interest. But they certainly do not know that they do.

The second familiar concept of political theory is even more abstract than liberty. This is equality, and in one way Rawlsian men and women cannot choose other than to protect it. The state of ignorance in the original position is so shaped that the antecedent interest of everyone must lie, as I said, in the same solution. The right of each man to be treated equally without regard to his person or character or tastes is enforced by the fact that no one else can secure a better position by virtue of being different in any such respect. In other contract situations, when ignorance is less complete, individuals who share the same goal may nevertheless have different antecedent interests. Even if two men value life above everything else, for example, the antecedent interest of the weaker might call for a state monopoly of force rather than some provision for private vengeance, but the antecedent interest of the stronger might not. Even if two men value political participation above all else, the knowledge that one's views are likely to be more unorthodox or unpopular than those of the other will suggest that his antecedent interest calls for different arrangements. In the original position no such discrimination of antecedent interests can be made.

It is true that, in two respects, the principles of justice that Rawls thinks men and women would choose in the original position may be said to fall short of an egalitarian ideal. First, they subordinate equality in material resources, when this is necessary, to liberty of political activity, by making the demands of the first principle prior to those of the second. Second, they do not take account of relative deprivation, because they justify any inequality when those worse off are better off than they would be, in absolute terms, without that inequality.

Rawls makes plain that these inequalities are required, not by some competing notion of liberty or some overriding goal, but by a more

---

[26] P. 61.

basic sense of equality itself. He accepts a distinction between what he calls two conceptions of equality:

> Some writers have distinguished between equality as it is invoked in connection with the distribution of certain goods, some of which will almost certainly give higher status or prestige to those who are more favored, and equality as it applies to the respect which is owed to persons irrespective of their social position. Equality of the first kind is defined by the second principle of justice . . . . But equality of the second kind is fundamental.[27]

We may describe a right to equality of the second kind, which Rawls says is fundamental, in this way. We might say that individuals have a right to equal concern and respect in the design and administration of the political institutions that govern them. This is a highly abstract right. Someone might argue, for example, that it is satisfied by political arrangements that provide equal opportunity for office and position on the basis of merit. Someone else might argue, to the contrary, that it is satisfied only by a system that guarantees absolute equality of income and status, without regard to merit. A third man might argue that equal concern and respect is provided by that system, whatever it is, that improves the average welfare of all citizens counting the welfare of each on the same scale. A fourth might argue, in the name of this fundamental equality, for the priority of liberty, and for the other apparent inequalities of Rawls's two principles.

The right to equal concern and respect, then, is more abstract than the standard conceptions of equality that distinguish different political theories. It permits arguments that this more basic right requires one or another of these conceptions as a derivative right or goal.

The original portion may now be seen as a device for testing these competing arguments. It supposes, reasonably, that political arrangements that do not display equal concern and respect are those that are established and administered by powerful men and women who, whether they recognize it or not, have more concern and respect for members of a particular class, or people with particular talents or ideals, than they have for others. It relies on this supposition in shaping the ignorance of the parties to the contract. Men who do not know to which class they belong cannot design institutions, consciously or unconsciously, to favor their own class. Men who have no idea of their own conception of the good cannot act to favor those who hold one ideal over those who hold another. The original position is well designed to en-

---

27 P. 511.

force the abstract right to equal concern and respect, which must be understood to be the fundamental concept of Rawls's deep theory.

If this is right, then Rawls must not use the original position to argue for this right in the same way that he uses it, for example, to argue for the rights to basic liberties embodied in the first principle. The text confirms that he does not. It is true that he once says that equality of respect is "defined" by the first principle of justice.[28] But he does not mean, and in any case he does not argue, that the parties choose to be respected equally in order to advance some more basic right or goal. On the contrary, the right to equal respect is not, on his account, a product of the contract, but a condition of admission to the original position. This right, he says, is "owed to human beings as moral persons," and follows from the moral personality that distinguishes humans from animals. It is possessed by all men who can give justice, and only such men can contract.[29] This is one right, therefore, that does not emerge from the contract, but is assumed, as the fundamental right must be, in its design.

Rawls is well aware that his argument for equality stands on a different footing from his argument for the other rights within his theory:

> Now of course none of this is literally argument. I have not set out the premises from which this conclusion follows, as I have tried to do, albeit not very rigorously, with the choice of conceptions of justice in the original position. Nor have I tried to prove that the characterization of the parties must be used as the basis of equality. Rather this interpretation seems to be the natural completion of justice as fairness.[30]

It is the "natural completion," that is to say, of the theory as a whole. It completes the theory by providing the fundamental assumption that charges the original position, and makes it an "intuitive notion" for developing and testing theories of justice.

We may therefore say that justice as fairness rests on the assumption of a natural right of all men and women to equality of concern and respect, a right they possess not by virtue of birth or characteristic or merit or excellence but simply as human beings with the capacity to make plans and give justice. Many readers will not be surprised by this conclusion, and it is, as I have said, reasonably clear from the text. It is an important conclusion, nevertheless, because some forms of criticism of the theory, already standard, ignore it. I shall close this long essay with one example.

---

28 *Id.*

29 Chapter 77.

30 P. 509.

One form of criticism has been expressed to me by many colleagues and students, particularly lawyers. They point out that the particular political institutions and arrangements that Rawls says men in the original position would choose are merely idealized forms of those now in force in the United States. They are the institutions, that is, of liberal constitutional democracy. The critics conclude that the fundamental assumptions of Rawls's theory must, therefore, be the assumptions of classical liberalism, however they define these, and that the original position, which appears to animate the theory, must somehow be an embodiment of these assumptions. Justice as fairness therefore seems to them, in its entirety, a particularly subtle rationalization of the political status quo, which may safely be disregarded by those who want to offer a more radical critique of the liberal tradition.

If I am right, this point of view is foolish, and those who take it lose an opportunity, rare for them, to submit their own political views to some form of philosophical examination. Rawls's most basic assumption is not that men have a right to certain liberties that Locke or Mill thought important, but that they have a right to equal respect and concern in the design of political institutions. This assumption may be contested in many ways. It will be denied by those who believe that some goal, like utility or the triumph of a class or the flowering of some conception of how men should live, is more fundamental than any individual right, including the right to equality. But it cannot be denied in the name of any more radical concept of equality, because none exists.

Rawls does argue that this fundamental right to equality requires a liberal constitution, and supports an idealized form of present economic and social structures. He argues, for example, that men in the original position would protect the basic liberties in the interest of their right to equality, once a certain level of material comfort has been reached, because they would understand that a threat to self-respect, which the basic liberties protect, is then the most serious threat to equal respect. He also argues that these men would accept the second principle in preference to material equality because they would understand that sacrifice out of envy for another is a form of subordination to him. These arguments may, of course, be wrong. I have certainly said nothing in their defense here. But the critics of liberalism now have the responsibility to show that they are wrong. They cannot say that Rawls's basic assumptions and attitudes are too far from their own to allow a confrontation.