CHAPTER X

Conclusion

43. THE SUBJECT OF SOCIAL JUSTICE

In this final chapter I shall draw on what has gone before to offer my answers to three questions about justice. The first is: What is justice? The second is: Why be just? and the third is: How do we go about determining what justice demands? I shall take these questions in turn, devoting this section to the first and taking up the others in the following two sections. I shall not confine myself to recapitulating points that have been made already in this volume. In the course of reordering and synthesizing the material from earlier chapters I have also taken the opportunity on occasion to extend or elaborate what I said the first time round.

To begin, then, let us ask: What is justice? The word "justice" is used in a wide variety of contexts. Perhaps the one that comes to mind most readily is justice as an attribute of individual legal decisions. At any rate as a first move we may say that a verdict in a trial is just if it is in accordance with the law. But suppose that the law on workers' compensation lays down limitations on recovery such that a worker is denied compensation if his disease becomes manifest only some years after he has ceased employment, even though it is not in contention that his disease was incurred in the course of employment. If we feel that the denial of such a worker's claim is unjust, while conceding that it was in accordance with the law, we must be saying that the law itself is unjust.

In this *Treatise on Social Justice* I am concerned with justice in its

wholesale rather than its retail form—with institutions rather than individual outcomes. Thus, the provisions of the law on workers' compensation falls within its scope. These provisions can be assessed for their justice or injustice, and my object is to provide a framework within which such judgments can be made. My subject is, then, social justice, or, as it is sometimes called, distributive justice. This kind of justice is in the first instance an attribute of institutions. We can say that an existing institution is just or unjust. We can say that some alternative to what exists would be more just. And we can say that it would be just for a kind of institution that does not now exist (for example, a scheme providing for systematic and nondiscretionary transfers of income from rich countries to poor ones) to be created.

Institutions may be assessed from many points of view. What, then, is the distinctive point of view of justice? When we ask about the justice of an institution we are inquiring into the way in which it distributes benefits and burdens. The currency of social or distributive justice is one of rights and disabilities, privileges and disadvantages, equal or unequal opportunities, power and dependency, wealth (which is a right to control the disposition of certain resources) and poverty. It should be apparent from this that the justice or injustice of an institution is an enormously important fact about it. The judgment that an institution is unjust must tell very strongly against its overall acceptability.

At the same time, however, we should recognize that to ask about the justice of an institution is to look at it in one particular light: it is to look at it as a creator of benefits and burdens. (I shall take "benefits and burdens" as the generic term for those factors listed above.) Thus, aspects of institutions other than the distribution of benefits and burdens that they bring about are left on one side. In cases where the distribution of benefits and burdens is incidental to the rationale of an institution, asking whether the institution is just or unjust may be somewhat beside the point.

An obvious example is that of the public subsidization of grand opera. There is no country in which the highest level of performance can be sustained financially by box office receipts, and the balance is made up out of some mix of private and public support. (In the United States the element of public subsidization comes about mainly through tax relief on charitable contributions and the favorable tax treatment of foundations. The effect on the taxpayer is the same: the instrumentality simply happens to be that taxes which would otherwise have been collected are forgone rather than that taxes are collected and then

disbursed.) Now, regarded as a scheme for the distribution of benefits and burdens, the public subsidization of grand opera is bound to seem rather bizarre. For looked at in this light what it amounts to is a scheme for burdening the body of taxpayers and benefiting those who like grand opera and can afford the (often still quite expensive) tickets—and also, of course, benefiting the singers, orchestral players, and others whose salaries are partly paid for by the public subsidy.

The argument against subsidization is, indeed, quite frequently made along these lines. It is an argument that appeals to those on both the philistine left and the philistine right. The former can argue that transfers should not be made to those who have above-average incomes (as most operagoers have), and the latter can argue that grand opera should be treated no differently from any other marketable commodity, so if it cannot be sold at a price that covers its cost it should not be provided at all. But the point of subsidizing grand opera is not to improve the equity of the distribution of benefits and burdens. These arguments do not therefore show that there is anything wrong with subsidization. The most they can show is that it would be crazy to make an argument in favor of subsidization by saying that it is a demand of social justice. But has anyone ever made the argument on these lines? The case must be, rather, that grand opera represents one of the achievements of Western civilization and it is good that some minute fraction of a wealthy country's national income should be devoted to keeping it in being for current and later generations. This is not the place to elaborate or evaluate that case. The point that is relevant here is that it is an example of a public policy issue to which justice is essentially irrelevant.

More common are cases where justice is very important but is not the only consideration that matters. An educational system provides a good example. There are, obviously, many ways in which a society's educational institutions may be evaluated. We may approve or disapprove of the attitude to life that the schools seek to inculcate, for example, and we may assess what is taught according to its truth, its significance for the understanding of the world, its contribution to economic production, and so on. But when we look at educational institutions from the point of view of justice, what we will tend to focus on is the role that they play in the transmission of occupational positions from generation to generation. For when we ask what impact the educational system has on the distribution of benefits and burdens, the feature of it that becomes prominent is the way in which educational qualifications are the key to access to many of the more desirable occupational positions. The

educational system is thus seen as a system for the more or less equal distribution of opportunities to acquire these qualifications.

From this point of view, the content of education is of far less significance than the method of allocating the chances to acquire whatever educational qualifications are most valuable. Let me illustrate this. Bureaucratic positions, in societies otherwise as different as ancient China and modern Britain, have been filled on the basis of performance on examinations whose content has little to do with the knowledge required for the carrying out of the duties. Macaulay defended the practice of recruiting to the Indian civil service on the basis of the ability to compose verses in Greek and Latin by arguing that this was as good a test as any of general ability. If it were made known that in future skill in Cherokee would be the criterion, that would do just as well: those who could produce the most mellifluous imitations of the best Cherokee models would undoubtedly be the best-qualified candidates for administrative positions.

One may reasonably question Macaulay's belief that general ability, as manifested in the ability to write elegantly in a foreign (and preferably dead) language, is the ideal basis on which to recruit a civil service. But that is not in itself a question of justice. What makes it a question of justice is that the content of the examinations has a profound influence on the access to the jobs that are filled by them. If (as was the case) Greek and Latin form three-quarters of the curriculum in schools available only to the sons (daughters at that time were not in the running) of the most privileged minority of the population, and are little taught in the rest of the schools, the examination system operates as a method of restricting recruitment to the children of that small stratum.

Thus, we look at the interaction of the educational system and the system of recruitment into desirable positions in the society, and we pass a judgment on the distribution of benefits and burdens that together they bring about. This gives us an assessment in terms of social justice. Clearly, the justice of all the other major social institutions can be appraised in the same way. It will be the task of volumes II and III of this *Treatise* to show how this is to be done.

44. JUSTICE AND MOTIVATION

Our account of the nature of justice cannot be separated from the question of motivation. What is the claim that justice has upon us? Or, as it is often put nowadays: Is it possible to show that it is rational to con-

form our conduct to the demands of justice? This question must be raised because an inquiry into justice is not a purely theoretical investigation akin to the development of scientific theories about natural phenomena. If we discover something new about muons or gluons this does not lead directly to any implications about what (if anything) we should do about it. But if we reach the conclusion that we are behaving unjustly, it seems that we should somehow be failing to recognize the significance of this if we were to say merely "How interesting!" Unless the acknowledgment of injustice has at any rate some tendency to lead to a determination to do something about it, there seems little point in even talking about justice. We cannot therefore separate the question "What is justice?" from the question "Why be just?"

Before I ask about the motivation for behaving justly, however, I must back up and deal with a prior question. I have claimed that social or distributive justice is a virtue of institutions. But if this is so, how did we somehow start talking about justice as a virtue of people? Presumably there must be some connection, but what is it?

It is extremely hard, in fact, to give a general answer to this question, but I shall offer a partial one that I hope will be adequate for the present purpose. The most straightforward case is one where an existing institution is just. We still, of course, have to say what it is about a distribution of benefits and burdens that makes it just. For now all we need is to take it that justice is a criterion for the assessment of institutions in terms of the distribution of benefits and burdens to which they give rise. Let us say, then, that we have a just institution in some society. Derivatively, we can say some things about just conduct in this instance. The central claim is that it is just to comply voluntarily with the requirements that the institution lays on individuals, at any rate so long as the institution is broadly complied with by those who are covered by it. Sanctions against noncompliance are relevant to justice insofar as they provide assurance that others will comply, so that one will not be put at a relative disadvantage by complying. But justice as a virtue of human beings is not exemplified by compliance that is brought about through fear of sanctions. Rather, justice is a disposition to do what just institutions call for simply on the basis of their being just.

This is all very well as far as it goes, but what does the virtue of justice require in the face of unjust institutions? Since there is a range of reasonable disagreement about what justice requires, no society could subsist for a day if people felt no obligation to comply with institutions that they believed to be unjust. We therefore need some quite robust

conditions to be met before justice as a virtue of individuals ceases to commend compliance. We have to look at the overall tendency of the society's institutions and the way in which political decisions get taken in order to form a judgment.

But what if, after we have done that, the injustice appears to be egregious? Should we do what would be required by more just institutions? Should we refuse to obey the existing rules? What means are legitimate in seeking to change them? These are questions of right conduct that are extraordinarily hard to say anything about on anything except a case-by-case basis. And even then I do not believe that political philosophy has a lot to contribute. I shall, however, attempt to address them in Volume III of this *Treatise*.

For the present, let us lay aside these further questions about justice as a virtue of individuals, and take the simplest case, where institutions are just. In this case, as we have seen, justice consists in voluntary compliance with the demands that the institution makes. We can now go back to the question with which we began: Why be just? In the history of Western speculation on this topic (at any rate in its secular forms) there have been two kinds of answer. *Theories of Justice* has been built around the elaboration and comparison of these two kinds of answer.

Because of the practical nature of justice, a theory of the motivation for being just must at the same time be a theory of what justice is. For the content of justice has to be such that people will have a reason for being just. Thus, suppose we say that the motivation for being just is a sense of the long-term advantageousness to oneself of being just. Then justice must consist in what everyone finds advantageous if it is to be something that everyone has a motive for pursuing voluntarily. Alternatively, suppose that the motivation for being just is the desire to act in ways that can be defended to oneself and others without appealing to personal advantage. Then justice will have to be whatever it is that can actually be defended in this way. These are the two approaches that I regard as the main candidates to be considered. Each is, as may be seen, at once a theory of motivation and a theory of what justice is. Let me say a bit more about each in turn.

The first approach to justice can be, and has historically been, arrived at from two directions. Among the ancient Greeks, it was commonly held that every virtue must be advantageous to its possessor. Justice, however, seems on the face of it to be advantageous to others rather than to the person who is just. The thought then arises that under the actual conditions of human life the self-restraint demanded by justice

may be able to be presented as advantageous when viewed as the price of similar self-restraint by others. A conception of justice on these lines is mooted in Plato's *Republic*, as I mentioned in section 1, and Socrates does not in rejecting it question the premise that if justice is to count among the virtues it must be advantageous to its possessor.

From the seventeenth century on, this first approach has tended to follow from a theory of human motivation in general. Philosophers have held either that human beings do in fact ineluctably pursue their own interests, or that even if they sometimes diverge from the path of self-interest the only rational course of action is that which advances the agent's interests. It is clear that, on these premises, if justice is to have the capacity to appeal, it cannot be advantageous to others at the expense of the agent. And, as before, the most plausible way of making justice fit the requirements is to suggest that justice really is advantageous to everyone. Of course, if justice is to have content, it will be necessary to define that content in such a way as to ensure that it really is true that justice is advantageous to everyone, and there is no guarantee that justice so defined will correspond to what is commonly thought of as justice.

It is important not to render this first approach in a way that makes it appear cruder than it really is. Thus, Hobbes, the most famous expositor of the modern version of the first approach, has frequently been interpreted as saying simply that "might makes right." But this is not so. What justice consists in is, roughly speaking, carrying out undertakings that you consented to out of a sense of their advantage, and Hobbes maintains that (under certain conditions that define the scope of obligation) it is advantageous to carry out such undertakings.[1] The neo-Hobbesian theory of David Gauthier is similar in this respect.[2]

I emphasize this point because I earlier defined the virtue of justice in people as a disposition to do what justice demands voluntarily, and not only under threat of sanctions. It is easy to assume that the Hobbes/Gauthier line of analysis cannot accommodate this because it has to make personal advantage the motive for cultivating the disposition. But not every source of advantage is properly to be regarded as a sanction. If we understand sanctions in a restrictive sense as evils formally and deliberately visited on people for lack of compliance, we can see that failures to be trustworthy may be disadvantageous in two ways that do not involve sanctions.

First, failure to do your part in some cooperative venture may cause it to collapse, or at any rate weaken it, so that you are eventually worse

off than you would otherwise have been. And, second, the cooperative scheme may continue after you fail to do your part but you may be excluded from this and perhaps other cooperative arrangements in future, to your long-run detriment. Hobbes makes both these points and also recognizes what the significance of formal sanctions is in this scheme. It is not that they are necessary to provide a reason for cooperating when others cooperate. Rather, what they do is to give greater assurance that others will in fact cooperate, thus ensuring that it will be advantageous to cooperate yourself.

It is worth getting clear what a sophisticated version of the first approach looks like so as to get the relation to the second approach straight. It is not the case that on the first approach anything that is advantageous for someone to do is just. Justice consists in playing one's part in mutually advantageous cooperative arrangements, where the standard of comparison is some state of affairs defined by absence of cooperation. Now in my view there is absolutely no question that this is at any rate a part of justice. Reciprocity is a core element in every society's conception of justice. The question is, rather, whether the first approach is adequate by itself. My answer is that it is not and that the second approach yields a general theory of justice which can incorporate the theory generated by the first approach as a special case. I shall explain at the end of the next section how this can occur, but I must now say something about the second approach itself.

It may be recalled that I defined the second approach more or less as the obverse of the first, saying that according to this approach the motivation for being just is the desire to act in ways that can be defended to oneself and others without appealing to personal advantage. Let me first follow along the line thus started by illustrating how this second approach differs from the first and then try to set out the foundational ideas underlying the second approach independently.

I have denied that the first approach is simply one of "might-makes-right." This would be to ignore the contractual basis of the theory. Nevertheless, the charge is not wholly misplaced. For might can be transformed into right by the alchemy of consent. Thus, Hobbes insists on the validity of any agreement that a weak state makes with a strong one as the price of peace, however disadvantageous its terms may be. And this implies that it would be unjust to go back on it if the opportunity arose to do so with impunity. (Hobbes does allow that a new and just cause of fear would invalidate the agreement, but the case I am considering is one where fear is subsequently lessened rather than

increased.)[3] Similarly, someone who is captured in a war and is spared his life in return for a pledge of servitude is obliged by this "covenant of obedience" to put his labor and that of his family at the disposal of the master.[4]

Gauthier, it should be said, attempts to ameliorate such results of the doctrine of justice as mutual advantage by ruling out threats as a way of creating the noncooperative alternative. As I argued in section 37, I do not think that he can do so compatibly with his premises. But in any cases he insists that natural advantages and disadvantages must be translated into unequal bargains by the method of splitting the difference, and that it is a disposition toward conformity with these bargains that constitutes the virtue of justice. (See especially sections 5 and 7.)

Now the alternative to this approach is to deny that the motivation for being just has to be its prospective advantageousness, and therefore to deny that the only basis of justice can be mutual advantage in comparison with the outcome if no agreement is reached. From this perspective, part of the point of justice is to provide a criterion for the redress of inequalities in bargaining power. Justice is not supposed to be merely a device for smoothing the path of exploitation, a way of ensuring that those with the stronger bargaining position are able to turn it automatically into an advantageous outcome.

How can we put the second approach positively rather than, as has been done so far, negatively? It is, I think, true to say that whenever someone wishes to deny that a distribution of benefits and burdens is just, while acknowledging that it is mutually advantageous to the parties in the situation as it actually exists, the same general kind of appeal is made. This is an appeal to what can be approved of from an impartial standpoint. Thus, if we call the first approach justice as mutual advantage we may call the second justice as impartiality.

The basic idea of justice as impartiality can be expressed in a variety of ways. One is the notion of an impartial observer: justice is seen as what someone with no stake in the outcome would approve of as a distribution of benefits and burdens. Another is to ask each of the parties, "How would you like to be treated in the way you are proposing to treat the other?" The object here is to get the party that stands to gain from an inequality of bargaining power to admit that it would not like to be on the losing end.

This basic idea of impartiality as a matter of putting oneself in the other's shoes can be fleshed out in a couple of different ways. One is to ask the parties what outcome they would favor if they did not know

which position they occupied. The idea of this is to guarantee impartiality by preventing any party from giving the answer that suits itself. The other is to ask the parties to propose principles for the distribution of benefits and burdens that they think ought to be acceptable to everyone affected not merely as preferable to the outcome arising from lack of agreement but under conditions in which that kind of bargaining pressure is removed.

Each of these variants of the second approach can be formalized further and thus be made to constitute a separate version of justice as impartiality. Indeed, even saying this greatly underestimates the variety of theories that are possible—and that actually exist. For we must allow that each variant can be formalized in different ways, and the details of these differences can produce profound effects on the outcomes that the theory generates as constituting a just distribution of benefits and burdens. I shall return to this in the next section, which is concerned with the question of determining what is just. For the present purpose, it is sufficient if I have said enough about justice as impartiality to give a foundation for the discussion of motivation that is the business of this section.

What, then, is the motive for behaving in conformity with justice, conceived of along the lines of the second approach? To provide an answer it is necessary to challenge the theory of motivation that underlies the first approach. Whether in the Greek or the modern form, what the first approach comes down to is the claim that, if something's being just is to count as a good reason for doing it, justice must be shown to be in the interest of the agent. On the second approach this constraint on what can count as a good reason is abandoned. That something is just, as justice is understood by the second approach, can be in itself a good reason for doing it. The motive is the desire to act justly: the wish to conduct oneself in ways that are capable of being defended impartially.

No doubt it is simpler to appeal to one motive rather than two. This may account for the popularity of the first approach among undergraduates who are beginning to study moral philosophy. But a theory may be too simple to be adequate. It seems to be quite well established that human beings are and always have been moved by considerations of justice as impartiality. If it is said that they are irrational to give weight to such considerations, what exactly is the force of this? Presumably it is not being claimed that they are acting on some kind of factual error. And to say that it is irrational because the only rational motive is self-

interest merely assumes what needs to be proved. At the same time, I do not wish to endorse the claim made by some philosophers that it is irrational not to act in accordance with justice as impartiality. What I am saying is that the desire to be able to justify our conduct in an impartial way is an original principle in human nature and one that develops under the normal conditions of human life.

It may be said that, even if it is not irrational to be disposed to act according to the dictates of justice as impartiality, it is irrational to act on the assumption that others will do so, and irrational to create institutions that rely upon their doing so. For practical purposes should we not address ourselves to the universal motive of self-interest rather than to the weak and undependable force of justice as impartiality?

This claim can be effectively undermined, in my view, by bringing to bear a number of mutually reinforcing counterarguments. Suppose, to being with, that we were to concede the greater reliability of self-interest as a motive. The chain of reasoning that purports to show the long-run advantage of, say, keeping one's agreements is quite complex. Even if we were to concede its validity, we should have to say that it is going to be hard to harness self-interest to justice in the way proposed by Hobbes and Gauthier. The essence of justice as impartiality is encapsulated in the Golden Rule; the efforts of Hobbes and Gauthier require hundreds of pages of subtle reasoning. But I do not think that even then the chain of reasoning leading to justice as mutual advantage is completely secure. It is not really possible to prove that it is advantageous to be disposed to be just on all occasions—for example, to adhere inflexibly to a policy of keeping contracts that are, according to the terms of the theory, fairly entered into. As I suggested in section 19, there is no answer to Hume's "sensible knave" except the one that Hume himself offers, and that appeals to the force of the moral motive rather than to self-interest.

The notion that impartial justice is a weaker motive than that of self-interest should not, in any case, go without challenge. To see its importance we have only to contrast a world in which everyone accepts that the only rational motive is self-interest with one in which it is accepted that it can be rational to do things in pursuit of justice that are contrary to one's interest. In the first world, bargaining advantage is smoothly translated into outcome advantage. Each party dispassionately appraises the relative bargaining strengths and settles for the best deal it can hope to get. Now add the factor of a sense of justice that is not simply a device for ratifying inequalities of bargaining power. "Instead of everyone's wanting as much as he can get in a bargaining situation,

suppose there is some outcome such that everyone cares very much about getting his share under that outcome, but cares very little, not at all, or even negatively about getting more. Suppose also that each person prefers carrying out his threat to settling for less than his share under that outcome."[5] Then we can expect that outcomes in the second world will be different from those in the first.

There is no need to overstate the case. So long as one party has a large enough advantage in bargaining strength, it will still tend to finish up with more than justice as impartiality would give it unless it actually disvalues ill-gotten gains. But the overall effect of the sense of justice will be to shift outcomes in a just direction. If they believe their cause is just, armies will be less willing to surrender than a rational calculation of advantage would suggest. Similarly, trade unionists will be more willing to strike if they believe they have a just claim than they would be in the world of dispassionate calculation of relative bargaining advantage.

It is worth noticing that the determination to accept losses in the pursuit of just claims over and above what would be suggested by self-interested rational calculation actually itself changes the relative bargaining strengths of the parties: the other side is more likely to settle for the demands if they are backed by moral fervor derived from a sense of justice. But even if the threat has to be carried out—the soldiers carry on fighting, the union strikes—the outcome may still, at some cost to the weaker party, be closer to the just one than it would be in the world without a sense of justice.*

I cannot imagine what it would mean to say that, in the aggregate, the sense of justice has (say) 20 percent of the efficacy of self-interest. What can be said is, surely, that it conditions a vast number of everyday transactions in the world we live in, and that, where the sense of injustice is deep and pervasive, it can give rise to deeds of heroic self-sacrifice that prospects of personal gain could scarcely do. It would be hard to explain the political behavior of Palestinians or South African blacks by

* For what it is worth, we can see a parallel in a stylized version of fighting for territory within, for example, many species of birds. Suppose that a bird fights harder the closer it is to the center of its territory. Then territories will be more equal than they would be if birds fought equally hard wherever they met, though it is consistent with the model that a stronger bird should have a larger territory than a weaker one. Moreover, if it is common knowledge among adjacent birds that the outcome of a fight depends on its location, fights will occur, if at all, only at the boundaries of territories. The sense of justice functions in the same way as the sense of territory postulated here: what we are saying is that equal increments of gain will elicit unequal amounts of prospective loss to obtain them, depending on their perceived justice.

assuming that individual self-interest is the sole motivation. And those (mostly to be found in the United States) who are so wedded to that assumption that they dismiss everything that does not fit as "irrational" are condemned never to understand the rest of the world.[6]

There is a further point so obvious that it is sometimes overlooked when we talk about the motivation to do what justice requires. We may say, as I have done, that justice as a virtue is realized only in the voluntary compliance with what justice requires. And I think that impartial justice would be hard to instantiate if nobody ever regarded the injustice of an action as a reason for refraining from it. But I have now added that one can also be moved by a sense of justice to make strenuous efforts to satisfy just claims, and this is another and very important way in which justice as impartiality can be efficacious as a force among human beings. But we must add that voluntary compliance and self-help do not have to stand alone. The basis of justice is institutional, I have argued, and institutions normally deploy sanctions to provide an additional motive for compliance. It is not, therefore, necessary that everyone should be moved by a sense of justice so long as the gap can be filled by deliberately created incentives for compliance.

This point is of especial importance because I think that the sense of justice often has more free play at the stage when the forms of institutions are being decided on than at the stage when people are deciding whether or not to comply with the demands that institutions make on them. Self-interest cannot be expected to bring about just institutions in general, so it is crucial that the sense of justice should operate there. Fortunately, all that is often necessary is that those whose own interests are not directly affected should support the course of impartial justice. But people who are prepared, say, to vote for a fair system for assessing contributions to some collective project may not be sufficiently motivated to pay their contributions voluntarily. The solution here is, obviously, to vote also for a system of sanctions to ensure compliance.

45. THE CONTENT OF JUSTICE

The third and last question that I raised about justice is: How do we determine what justice requires? The answer will, naturally, depend on which of the two conceptions of justice we accept. Let me first ask how we can get from the general idea of justice as mutual advantage to some determinate conclusions about just institutions. In section 33 I presented it as a contractualist theory. The paradigm of mutual advantage

is a contract: an agreement to move from some *status quo* to some new arrangement that is prospectively beneficial to both parties. Now it is an obvious and familiar objection to any kind of contractarian doctrine that institutions do not in fact rest upon agreements among millions of people, nor is it plausible that they should. And the response among sophisticated contract theorists has always been the same: that the social contract, unlike particular contracts, must be regarded as hypothetical. Thus, for Hobbes all specific social institutions are created and sustained by the fiat of the sovereign, but the sovereign's authority rests on a hypothetical contract: a "Covenant of every man, in such a manner, *as if* every man should say to every man" that he will give up his rights in favor of a sovereign provided everybody else does likewise.[7] And his disciple Gauthier similarly has social cooperation depend upon hypothetical rather than actual agreements.

This response is normally countered by a further objection, namely, that it is hard to see why hypothetical contracts should have any binding force. If I enter into a real contract, it is my voluntary decision that gives the contract its moral claim on me. Saying that it would have been advantageous to me to make some contract that I did not in fact make establishes no similar claim on me.

This objection would have a good deal of force against many versions of social contract theory, but it fails against the Hobbes/Gauthier version. For we must always recall that its initial premise is that the binding force of agreements rests in the sense of mutual advantage that presumptively underlies them. An actual agreement signalizes a recognition of mutual advantage but it cannot create a motive *de novo*; all it can do is channel the motive of self-interest. Justice thus underwrites mutually advantageous cooperative arrangements, whether they arise from explicit agreement or not. If the argument from long-term self-interest works in the one case, then it works as well in the other case. Cooperation is preferable to noncooperation—that is the central point. It is immaterial whether the cooperation arises from an actual contract or whether its basis lies in a convention that each should support cooperative arrangements on condition that others do the same.[8]

The question we are now confronted with is the following: given that we cannot appeal to actual agreements in relation to social institutions, how are we to establish what are the just terms of cooperation? The natural answer would seem to be that the just terms of cooperation are those that would have been agreed upon by people trying to do the best for themselves. For these are, presumably, precisely the terms that it is

rational for self-interested people to support amd maintain here and now.

In order to give effect to this basic idea we need two things: a nonagreement point from which the hypothetical bargaining is to start, and a theory about the outcome of bargaining among rational self-interested agents. Thus, sticking to Hobbes as our example for a moment longer, we can say that he has an explicit answer to the first and an implicit answer to the second. His noncooperative baseline from which mutual advantage is to be reckoned is "the natural condition of mankind," which is a state of war where "the life of man" is, in the famous phrase, "solitary, poore, nasty, brutish, and short."[9] And his implicit answer to the question about the rational terms of cooperation is that any improvement over the nonagreement baseline is enough to warrant consent to the terms of cooperation. Thus, he admits that "of so unlimited a power" as he ascribes to the sovereign, "men may fancy many evil consequences." But he replies that "the consequences of the want of it, which is perpetuall warre of every man against his neighbour," are much worse.[10]

It is apparent that Hobbes's way of specifying the noncooperative outcome and his criterion for rational cooperation have the combined result that the range of just outcomes is extremely large. This is in fact Hobbes's intention: it is rational, according to Hobbes, to obey the sovereign unless doing so would constitute a direct threat to one's own life. If we want to get tighter constraints than these out of the notion of justice as mutual advantage we shall have to make the nonagreement point less abysmal and put more teeth into the criterion for rational bargaining outcomes.

Gauthier, as it happens, illustrates just how far the opposite position can be carried. On essentially Hobbesian psychological premises he erects a wonderful Lockean farrago which has fully developed market institutions all built into the noncooperative outcome. Overlooking Adam Smith's remark that, left to themselves, profit-seeking producers will collude to stifle competition, Gauthier supposes that a competitive equilibrium could in principle arise purely from the operations of self-interest in a state of nature. Thus, the only room left for cooperation is to deal with externalities and public goods. And here the second move to tighten up the constraints of justice on social institutions comes into play. It is not enough that everyone should be better off than at the nonagreement point. Rather, there is, according to Gauthier, a unique solution to every bargaining problem, and an institution is just only if it corresponds to that solution.

What kind of theory do we get when we try to operationalize the notion of justice as mutual advantage so as to elicit from it definite (and preferably unique) implications for the justice of institutions? I answer that what we get is a variety of constructivist theory. A constructivist theory (see above, section 33) tells us that, in order to get answers about the demands of justice, we must construct a model of human interactions in some specified context. The context includes the motivations of the actors and the rules of the game. The theory then says that what comes out of the hypothetical interactions is to be taken as constituting principles of justice. In the present instance, we have self-interest as the motive. We have some stipulations about the kinds of moves that can be made in establishing the nonagreement point. (For example, Hobbes admits but Gauthier excludes threats.) And we have some formula that tells us how rational self-interested actors move from the nonagreement point to the Pareto frontier—that is to say, how the gains from cooperation are to be divided up.

What, then, is to be said about the operationalization of the second approach, justice as impartiality? To answer this I can draw on the analysis of constructivist theories already sketched. For it will hardly have escaped notice that all the formulations of the notion of justice as impartiality that I ran through in the previous section included hypothetical elements: "How would you like it if. . .?" "What would you say if. . .?" and so on. And if we want to take such vague notions and generate definite implications from them about the requirements that must be met by just institutions, a constructivist theory of some sort seems like a natural recourse.

As before, a constructivist theory can be defined by the context that it sets up for the interactions of the hypothetical parties. As before, we are to identify what is agreed upon by the parties as the principles for just institutions. And, as before, we may leave it open whether the construction generates unique outcomes or only a range of acceptable outcomes. What is distinctive about the second approach is that a different set of considerations goes into the specification of the context.

Given the motivational assumptions of the first approach, it is clear that the construction must represent bargaining advantages and disadvantages; otherwise the principles derived from it would fail to elicit the allegiance of everyone. The second approach is distinguished from the first precisely by its denial of the assumption that people (or people insofar as they are rational) are moved only by self-interest. It rejects the idea that a theory of justice can have efficacy only to the extent that it makes the principles of justice reflect actual power relationships. It will

therefore throw out the nonagreement point of a "state of nature" as a generally relevant place at which to start. It may have room for nonagreement points in a subordinate capacity, but these will be derived from the theory itself, rather than being imposed on it. And it will in one way or another block the translation of superior bargaining power into advantageous outcomes.

How can the context of interaction be specified so as to meet these desiderata? There are, broadly speaking, two routes that can be taken. The first is to retain from the first approach the postulate that the parties in the construction are pursuing their own advantage but to prevent them from abusing superior bargaining power by denying them various kinds of knowledge—most crucially, knowledge of their own identities. The other is to drop the postulate that the parties are to do the best for themselves that they can and instead to postulate that they are, under ideal hypothetical conditions, seeking to reach agreement on principles that nobody could reasonably reject.

The postulate of self-interest plays quite a different role when it is incorporated into a construction designed to formalize the second approach from that which it played earlier. It is no longer intended to represent people as they really are—the essence of the second approach is that people can be concerned with justice as impartiality—but is simply part of the context of interaction from which, it is claimed, justice as impartiality will arise. We might say, however, that the two constructions are similar in that they take agreements arising out of the pursuit of self-interest in suitably defined conditions to be constitutive of justice. Where they differ is in the specifications of the conditions.

But it is not enough to show that a veil of ignorance superimposed on self-interested choice will guarantee impartiality of a kind. We can still ask: Is it the right kind? What was said about the force of hypothetical agreements within the first approach applies also, *mutatis mutandis*, to the second approach as well. If we take justice to be a matter of mutual advantage, the only claim to be made for hypothetical agreements is that they function as a sign of where mutual advantage lies. Similarly, if we take justice to be a matter of impartial approval, the only claim to be made for hypothetical agreements is that they are a sign of what might be approved of impartially. As I argued in section 41, however, self-interested choices under uncertainty fail to capture the notion of justice as impartiality that originally attracted us.

The construction that I have been discussing starts by retaining the self-interest postulate from the construction corresponding to the first approach, and then adds a veil of ignorance with the object of mitigating its undesirable effects. The alternative route is to throw out the idea that the "agents of construction" should be assumed to be trying to maximize their gains. Instead we postulate that they are trying to reach agreement on terms that nobody could reasonably reject. The veil of ignorance that prevents them from knowing their own personal characteristics and positions is then no longer essential. It may still be useful as an *ad hoc* device: sometimes an argument can be advanced by asking, "What would you say about this if you didn't know how you'd be affected?" But this is now a part of the arsenal of persuasion rather than a move in a knockdown demonstration.

I have identified three kinds of construction, one corresponding to the first approach and two to the second approach. Each embodies a different decision process. The first construction creates a game: we ask what rational self-interested players would finish up with. As Anatol Rapoport put it in *Fights, Games, and Debates*, the parties in a game "co-operate in 'doing their best,' that is, in presenting to each other the greatest possible challenge. . . . [T]he *assumption* that the opponent will do 'his best' contributes to the validity of rational analysis, which both must accept."[11] The second construction, involving a veil of ignorance, has superficially the same game-like characteristics in that the parties are pursuing their own interests as effectively as possible. Moreover, they know that in real life they have conflicting interests. But the proviso that they do not know what these interests are prevents them from being able to advance their distinctive interests in the choice situation—the "original position," as Rawls dubs it. Each therefore in practice faces the same decision problem in an original position so defined. But with no conflict of interest in the original position we lose the characteristics of a game. What we have instead is a problem of individual choice-making under uncertainty which is posed in identical terms to the people in the original position. Finally, the third setting is one that Rapoport calls, in contrast to a game, a debate. "The objective is to *convince* your opponent, to make him see things as you see them."[12] But we should add that conviction must be conceived of as a two-way process. The parties in this third construction must debate in good faith, which means that they must be prepared to be convinced as well as to try to convince others. They must be willing to acknowledge a good argument, even if it runs against their interests to do so.

So far I have specified the third construction in two ways. The parties are concerned to reach agreement on principles that nobody can reasonably reject; and they do not operate behind a veil of ignorance, though they may invoke it in the course of argument. We should add that the

parties are able to understand the implications of alternative proposals and are aware of a wide range of cultural and historical experiences. For some purposes it may help to think of the parties as representatives of the people in actual societies, but we should endeavor to avoid getting too hung up on details of this kind, since no claim is being made that conclusions can be demonstrated by using this apparatus. It is simply put forward as a way of thinking about justice as impartiality.

How does the construction connect up with the notion of impartiality? My answer is that impartiality enters in through the requirement that everybody's point of view must be taken into account. Each person in this original position has a veto over proposed principles, which can be exercised unless it would be reasonable for that person to accept a principle. To say that a principle could not reasonably be rejected by anyone covered by it is, I suggest, a way of saying that it meets the test of impartiality.

It has to be said that this construction is delicately poised between two poles, and that its integrity depends on maintaining an intermediate position. If it is pulled too far in either direction it turns into something else. On one side, we want to insist that the parties have interests and values that they are concerned, up to a point, to defend. But we do not want this to reduce the third construction to the first, where the parties utilize whatever strategic advantages they have in order to advance their interests. On the other side, we want to say that the parties are prepared to accept that it would be unreasonable to hold out against some proposal merely because it is relatively disadvantageous. But we do not want to say that their sense of what is reasonable is so strong that it leads them directly to identical conclusions about what is just. If we did this, we would lose the character of a debate, as the second construction lost the character of a game. Instead of a debate, we would again have an individual decision problem carried on in identical terms by all parties, and the requirement of consensus would once again be reduced to triviality. The only difference would be in the formulation of the decision problem: rather than an attempt to do the best for oneself under uncertainty it would now be an attempt to decide what justice requires.

46. ENVOI

"The Adventure of the Engineer's Thumb" is not one of the more distinguished items in the Holmesian canon, but it contains an incident that some readers of this book may regard as relevant. The engineer,

having been engaged by a gang of forgers who are anxious to conceal their location, is met at a country railway station and driven for an hour through the night in a carriage with closed windows, finishing up at a house which is actually quite close to the starting point.[13] It may perhaps be felt that I have similarly taken a long way round to arrive at a destination which is scarcely exotic. People who have not been exposed to the arguments of a Hobbes or a Gauthier will naturally tend, I think, to reject out of hand the notion that justice is nothing more than a matter of mutual advantage. And it may be added that few of those who have studied the arguments have been persuaded either. It is true that the exact specification of justice as impartiality proposed here would not naturally occur to someone who had not given the question much thought. But at the same time I have myself emphasized the way in which the general idea of justice as impartiality is a systematization of everyday forms of moral argument.

I do not, however, think that it is sensible to demand originality in the general nature of the conclusions of a study such as this. To return to what was said in the opening section of this book, questions about the justice of institutions arise when the authority of custom weakens its hold on the minds of the members of a society. As it comes to be perceived that social, political, and economic inequalities are the product of human convention, the need for justification is felt. Two responses that have been developed deny that social inequality is based on convention. One seeks to found social inequalities in natural ones, a line of argument that can be traced from Aristotle's defense of slavery to modern "scientific" racism. The other seeks a metaphysical basis for social inequalities: from the elaborations of the Hindu system to the Church of England's complacent belief that God "ordered" the "estate" of rich and poor, the major religions have a remarkable record of supporting whatever system of inequalities happens to prevail at the time. This book has been premised on the assumption that neither of these forms of justification for inequality will do. Although some people believe them, of course, they are incapable of carrying any rational conviction to those who do not. They cannot therefore form an acceptable basis for the justification of inequality.

It may be said that this criterion itself rigs the question by stipulating that only certain kinds of doctrine can constitute a legitimate basis for inequality. I can see no alternative to pleading guilty on this charge. There is, I think, an inevitable circularity here. Asserting the demands of unbelievers to be given better reasons for accepting inequalities from

which they suffer than those offered by the believers is necessarily denying the claims that the believers themselves make to the effect that their beliefs provide all the justification that can be required.

Many philosophers think that it is possible to resort to higher-level arguments to show the superiority of the view that justifications of inequality should not appeal to inherently controversial empirical or metaphysical beliefs. I wish they were right but I can see no way of overcoming the problem that they must attribute to their adversaries premises that they manifestly do not believe in. Thus, suppose we were to expand Locke's argument for religious toleration and say that there are strong pragmatic grounds for not trying to impose a social order based on a particular set of religious beliefs. Even if this is a sound case in itself, its persuasive power can be no greater than the appeal of pragmatic considerations to the recipient. Any true believer prepared to regard them as decisive is already in effect a convert to the premises of secularism in politics.

For the purposes of this *Treatise*, and any other on the same subject, it is necessary at the outset for the author to commit himself to a position on the range of justifications to be seriously considered. My commitment is manifest from everything in this book. The point that now falls to be made is simply this: that if we set the problem up as one of justifying inequality on the assumption that it is the product of human convention and not underwritten by any deep natural or metaphysical inequality between human beings, there are not a lot of potential solutions. At the highest level of generality, there are perhaps only the two that have been discussed in this book. Both start from the idea that conventions that are adhered to are preferable to unrestrained conflict. They then diverge in what they ask of a satisfactory convention. One line says that the convention must be acceptable to each person when he consults his own advantage. The other says that it must be acceptable to everyone when he takes up an impartial standpoint. Leaping lightly over the centuries, we can trace the first tradition from the Sophists through Hobbes to Gauthier and the second from the Stoics through Kant to Rawls.

When the alternatives are stated in terms such as these, it is easy to maintain that philosophy never makes any progress and that the same basic ideas merely keep coming round in cycles. There is obviously something in this, but I believe that it conceals more than it reveals. Progress comes in the form of analytical techniques that enable us to state the grand alternatives with greater precision than before and to see

more deeply into them. For the first alternative, the key development has been the invention of game theory and its increasingly flexible deployment in social analysis. For the second alternative, it has been the notion, put forward by Rawls, of an original position, conceived of as an ethically privileged choice situation. That these are genuine advances and not just changes of fashion is, I believe, evidenced by the way in which we can use these ideas to go back and shed fresh light on earlier versions of the theories. Thus, the level of sophistication with which Hobbes is treated has been raised immeasurably in the past two decades, and we are also beginning to see a reevaluation of the social contract tradition in the light of Rawls's work.

This book is offered as an attempt to consolidate and, more ambitiously, to carry further forward these advances of recent decades. I should like to believe that I have clarified a number of questions in the game-theoretical analysis of bargaining and have shown their relevance to the theory of justice as mutual advantage. I also hope that I have succeeded in following through on Rawls's passing suggestion that one might be able to construct a taxonomy of original positions within which Rawls's own version could be located as a special case. At the same time, I have tried to use the contrast between justice as mutual advantage and justice as impartiality as a way of probing the theories of justice put forward by Hume and Rawls. I do not want to make exaggerated claims for this exercise. No doubt by bringing some questions into sharp focus I have blurred others. But I feel confident of my basic claim that both approaches are implicit in the theories of both Hume and Rawls and I think it is illuminating to tease out the two approaches from their writings on justice.

In the end, however, it is true enough that the overall conclusions of the book can be stated in reasonably brief compass. The first three sections of this chapter are, indeed, designed to do precisely that. For the larger purposes of this *Treatise*, it would be very inconvenient if it were otherwise, since I wish in the two volumes that follow the present one to build on its conclusions. In the next volume, then, I shall take justice as impartiality to be my starting point and I shall then try to work through a number of problems that arise in its interpretation and application. In the third volume, I shall carry over the results of the second to address in detail one question: the application of the concept of social justice to the distribution of income and wealth.

15. Ibid., p. 525.
16. Ibid.
17. Ibid., p. 527.
18. Ibid., p. 569.
19. Gilbert Harman, *The Nature of Morality: An Introduction to Ethics* (New York: Oxford University Press, 1977), p. 162.
20. Ibid., p. 111 (italics in original).
21. Ibid., pp. 111–12.
22. Rawls, *A Theory of Justice*, pp. 289–90.
23. T. M. Scanlon, "Contractualism and Utilitarianism," in Amartya Sen and Bernard Williams, eds., *Utilitarianism and Beyond* (Cambridge: Cambridge University Press, 1982), pp. 103–28. The quotation is from p. 116 n. 12.
24. Thomas C. Schelling, *The Strategy of Conflict* (Cambridge, Mass.: Harvard University Press, 1960), chapter 4, pp. 83–118. See also David Lewis, *Convention* (Cambridge, Mass.: Harvard University Press, 1969).
25. See, for a discussion of this idea, Karol Sołtan, *The Causal Theory of Justice* (Berkeley and Los Angeles: University of California Press, 1987).
26. David Hume, *A Treatise of Human Nature*, ed. L. A. Selby-Bigge, 2d ed., ed. P. H. Nidditch (Oxford: Clarendon Press, 1978), bk. 3, section 3, pp. 501–13.
27. David Hume, *An Enquiry Concerning the Principles of Morals* in *Enquiries Concerning Human Understanding and Concerning the Principles of Morals*, ed. L. A. Selby-Bigge, 3d ed., ed. P. H. Nidditch (Oxford: Clarendon Press, 1975), pp. 195–96.
28. Rawls, *A Theory of Justice*, p. 584.
29. Ibid.
30. Ibid., p. 585.
31. Ibid.
32. Colin Turnbull, *The Mountain People* (New York: Simon and Schuster, 1972).
33. J. H. Wellbank, Denis Snook, and David T. Mason, *John Rawls and His Critics: An Annotated Bibliography* (London: Garland, 1982), p. 3, item A2.

CHAPTER 10

1. See Brian Barry, "Warrender and His Critics," *Philosophy* 48 (1968): 117–37; reprinted in Maurice Cranston and Richard S. Peters, eds., *Hobbes and Rousseau* (Garden City, N.Y.: Doubleday, 1972).
2. David Gauthier, *Morals by Agreement* (Oxford: Clarendon Press, 1986).
3. The passage runs as follows: "And if a weaker Prince, make a disadvantageous peace with a stronger, for feare; he is bound to keep it; unlesse (as hath been sayd before) there ariseth some new, and just cause of feare, to renew the war" (Thomas Hobbes, *Leviathan*, ed. with an introduction by C. B. Macpherson [Harmondsworth, England: Penguin, 1968], chap. 14, p. 198).
4. Ibid., chap. 20, p. 256.
5. Allan Gibbard, "Human Evolution and the Sense of Justice," in Peter A.

French, Theodore E. Uehling, Jr., and Howard K. Wettstein, eds., *Midwest Studies in Philosophy*. vol. 7: *Social and Political Philosophy* (Minneapolis: University of Minnesota Press, 1982), pp. 31–46. The quotation is from p. 37. I should point out that Gibbard's conception of a sense of justice is, in my terms, an emotional attachment to a bargaining solution.

6. The belief that everyone can be counted on to respond mechanically to a balance of advantage and disadvantage led to many of the barbarities of American policy in the Vietnam war such as the massive bombing of North Vietnam and Cambodia. See Alexander George and Richard Smoke, *Deterrence in American Foreign Policy: Theory and Practice* (New York: Columbia University Press, 1974).

7. Hobbes, chap. 17, p. 227 (my italics).

8. See, for the notion of contract by convention, Russell Hardin, *Collective Action* (Baltimore: Johns Hopkins University Press, 1982).

9. Hobbes, chap. 13, p. 186.

10. Ibid., chap. 20, p. 260. Compare the similar passage in chap. 18, pp. 238–39.

11. Anatol Rapoport, *Fights, Games, and Debates* (Ann Arbor: University of Michigan Press, 1961), p. 9.

12. Ibid., p. 12.

13. Sir Arthur Conan Doyle, "The Adventure of the Engineer's Thumb," in William S. Baring-Gould, ed., *The Annotated Sherlock Holmes* (New York: Clarkson N. Potter, 1967), vol. 2, pp. 209–24.

APPENDIX A

1. Based on diagram V (p. 29) of R. B. Braithwaite, *Theory of Games as a Tool for the Moral Philosopher* (Cambridge: Cambridge University Press, 1955).

2. Ibid., pp. 34–35.

3. Ibid., p. 34.

4. Ibid., p. 30.

5. Ibid., pp. 26, 27.

6. Ibid., p. 27.

7. Ibid., p. 28.

8. Ibid., p. 30.

9. Ibid., pp. 30–31.

10. Ibid., p. 33.

11. Ibid., p. 35.

12. Ibid.

13. Ibid., p. 32.

14. Ibid., p. 33.

15. Ibid., pp. 32–33.

16. Ibid., pp. 35–36.

17. Ibid., p. 32.

18. Ibid., p. 39.

19. Ibid., p. 40.

University of California Press
Berkeley and Los Angeles, California

Library of Congress Cataloging-in-Publication Data

Barry, Brian M.
Theories of justice.

(A Treatise on social justice; v. 1) (California series on social choice and political economy; 16)
Bibliography: p.
Includes index.
1. Social justice. I. Title. II. Series.
III. Series: Barry, Brian M. Treatise on social justice; v. 1.
JC578.B37 vol. 1 320'.01'1 s [320'.01'1] 88-27764
ISBN 0-520-03866-5 (cloth)
ISBN 0-520-07649-4 (ppb.)

Printed in the United States of America

2 3 4 5 6 7 8 9