Chapter 8

The Theory of Justice as Fairness

I

In the mid-twentieth century, John Rawls began work on a set of questions that quickly led him to formulate the central ideas of a new theory of social justice. Working steadily through the 1950s and 1960s, his labors led to the publication of *A Theory of Justice* in 1971. This long and intricately argued work, parts of which Rawls had circulated among scholars in the years leading up to its completion, had an immediate and major impact on academic political philosophy and beyond, stimulating a range of questions and inquiries that was far more extensive than that generated by any other theory of social justice in the twentieth century. Rawls called his theory "justice as fairness." The development and later elaboration of this theory occupied him for his entire professional life, from his first published essay in 1951 to his final efforts in 2000, just two years before his death. As one might expect, Rawls's thinking evolved over the nearly half a century he devoted to this work, with a particularly significant break in his conception of the theory occurring in the 1980s. In the short space I have available I shall for the most part ignore these developments, to focus on central features that remained relatively constant in the various statements of the theory.

A Brief History of Justice, First Edition. David Johnston.

Although Rawls was aware of the constellation of ideas about social justice that focused on the concepts of desert and need, the principal target of his criticism was utilitarianism, which in his view had come to dominate discussion of social institutions and policies so thoroughly as to exclude from serious consideration any alternative ways of thinking about them. Rawls offered several complaints about utilitarian theories. First, he argued that utilitarianism offers inadequate protection for liberty. Under some circumstances, it might be the case that happiness for a majority can best be attained by depriving a minority of persons of their liberty. If the aggregate gains in happiness to the majority are greater than the loss of happiness suffered by the minority, then the greatest happiness principle would justify the minority's loss of liberty. For Rawls, this possibility is sufficient by itself to demonstrate the inadequacy of the greatest happiness principle.

Just how likely this scenario is, is a question worthy of some debate. But it is at least a plausible scenario, and in rejecting utilitarianism Rawls had in mind momentous historical facts, as well as theories. Throughout his adult life, Rawls was profoundly conscious of the deep injustice that had been perpetrated by Americans of European origin through the enslavement of Africans and their descendants over multiple generations. Whenever he visited Washington, DC, he made a point of visiting the Lincoln Memorial, in recognition of the depravity of this practice and of the importance of its abolition. For him, any idea of justice that provides inadequate protection for liberty is necessarily flawed.

Rawls also argued that utilitarianism is based on a monistic conception of the good. What he had in mind here is that, by treating happiness as the sole ultimate measure of human well-being, utilitarian theory fails to accord due recognition to the fact that human beings have diverse interests and pursue diverse ends, of which happiness may be only one. On this point Rawls's view is a near relation of Kant's claim that human freedom rather than happiness should be at the focus of our ideas about justice. For

Rawls, it is an important, indeed fundamental fact that human beings embrace a variety of (what he called) conceptions of the good. Some people may believe that a life of happiness is the best kind of life a human being can have and that, ultimately, all other ends or objectives of life should be subordinated to the objective of attaining happiness. Others may consider a life of integrity in accordance with some particular conception of that virtue to be the best possible kind of human life, even if it must be purchased at the cost of happiness. Still others may hold still different ideas about the proper ends or objectives of human life. Rawls believed that utilitarianism does not take into account the full variety of human ends (or conceptions of the good), thereby failing to accord due recognition to the distinctive human capacity freely to formulate and to embrace a "plurality" (as Rawls and many other recent writers have called it) of legitimate conceptions of the good.

This criticism of utilitarianism may not be fully justified. Rawls himself seems to recognize that it may not apply to all the forms of utilitarian theory, and he accordingly defines the central object of his criticism as "classical" utilitarian theory, to which he believes Bentham, Mill, and Sidgwick subscribed. It is reasonable, however, to question Rawls's claim, even considering it to be directed only at these theorists. As we have seen, Bentham recognized and attempted to accommodate within the scope of utilitarian theory the fact that human beings hold "idiosyncratical" values; John Stuart Mill did the same. At least some of the utilitarian writers may be less vulnerable to this criticism than Rawls believed.

More generally, Rawls was dissatisfied with utilitarianism because that body of theory does not treat distributive questions as the central questions that must be asked about justice. In fact, generally speaking, utilitarian theories focus on aggregate human well-being, not on justice. Any claims these theories make about justice generally are derivative from and subordinate to claims about aggregate utility. In contrast, Rawls argued that questions about justice are the most important questions we can ask about social institutions. He declares,

on the opening page of *A Theory of Justice*, that "[j]ustice is the first virtue of social institutions [. . .] laws and institutions no matter how efficient and well-arranged must be reformed or abolished if they are unjust" (3/3). The terms "efficient" and "well-arranged" allude to the utilitarian values whose primacy Rawls was attempting to challenge. He often expressed the difference between his theory and its principal rival by asserting that, whereas the central concept in utilitarian theories is the concept of the good, from which the idea of the right must be derived, in the theory of justice as fairness the right is prior to the good (31/28).

In view of the status of utilitarianism as the prime target of his criticism, it is noteworthy that in one of his earliest essays, "Two concepts of rules" (1955), Rawls actually mounts a limited defense of utilitarianism. As we have seen, one familiar objection to utilitarianism is that the arguments utilitarians use to justify the punishment of wrongdoers could also be used to justify the "punishment" of innocent persons, if that practice would contribute to the good of society. Rawls argues that this criticism is misplaced.

Although Rawls offers a limited defense of utilitarianism against some criticisms in this essay, its main objects, as the title suggests, are to distinguish between two levels of argument about rules and to show that failure to observe this distinction has contributed to confusion in moral argument. He argues that there is a crucial difference between the justification of a *practice* and the justification of *actions within* that practice – and in this case, specifically, between justifying the practice of punishment and justifying actions within that practice. The practice of punishment can be justified (perhaps) by appealing to the greatest happiness principle. Actions within that practice, however, can be justified only by appealing to the rules by which that practice is constituted, not by appealing to the greatest happiness principle directly. The rules by which the practice of punishment is constituted are retributive rules, not (directly) utilitarian ones. So, according to Rawls in this essay, utilitarianism is not vulnerable to the criticism that it might justify the punishment of innocent persons. This misplaced

criticism (he argues) inappropriately applies utilitarian standards, which should be applied to the practice of punishment, to specific acts of punishment, which should be judged by retributivist standards, not by utilitarian ones.

Although Rawls's defense of utilitarianism against this criticism is not wholly persuasive, his essay is still of considerable interest for two main reasons. First, it reveals substantive inclinations that are confirmed throughout Rawls's writings from this point forward. Rawls took utilitarianism with the utmost seriousness. He considered it to be the pre-eminent theory for assessing social institutions and policies in his time and regarded it, consistently, as the most serious contender for this role – with the exception of his own theory of justice as fairness. He did not accord the same respect to retributivism. In the essay, he discusses utilitarianism and retributivism as *prima facie* rivals. Rawls suggests – with little argument – that retributivism cannot justify the practice of punishment. He argues for a division of labor between the utilitarian and retributivist views: the utilitarian argument, he suggests, provides a basis on which it might be possible to justify the practice of punishment, while the retributivist argument justifies actions within that practice. According to this conception of the field of argument, we can think of utilitarianism as a description of the point of view that a legislator should take up in considering whether to adopt the rules that constitute the practice of punishment, while the judge who is charged with applying those rules appeals to the retributivist point of view. It follows that "the utilitarian view is more fundamental." Even within its apparently "natural" domain, then, Rawls dismisses retributivism – a view of justice in relation to wrongdoing that is based on the concept of reciprocity – as a derivative and secondary view, one that has its proper place, but only in subordination to (what he believed to be) a more comprehensive and more fundamental theory. Rawls never seriously defended the assumptions that led to this dismissal.

A second interesting point about Rawls's essay is that it exemplifies a strategy he would deploy consistently throughout his career. He does

not simply reject retributivism. Instead, he subordinates it to an allegedly broader view, in this case the utilitarian view. For Rawls, retributivism is a valid approach to punishment, but only from within a point of view that is highly constrained, as the point of view of the judge is supposed to be constrained by the legislation he or she is charged with enforcing. This strategy of argument, which is strikingly reminiscent of the method that Hegel had deployed in his major philosophical works, from his *Phenomemology of Spirit* (1806) onward, contributed significantly to the magisterial impression created by Rawls's work and served as a kind of template, to which he would return repeatedly in contending with views that appeared to be at odds with his own.

II

Rawls describes the subject of his theory as the "basic structure" of society. A society's basic structure comprises its major social institutions, including its political constitution, its fundamental economic structures, and its principal social arrangements. For example, the institutions of private property in the means of production and of competitive markets are central components in the economic structures of some societies, whereas others have been based on collective ownership of the means of production and on command economies. Some countries' political constitutions provide strong legal protections for freedom of thought and for liberty of conscience; others do not. The monogamous family is a bedrock social institution in many societies, while in others the polygamous family in one form or another has stood for centuries as one of society's principal social arrangements.

What does the basic structure of a society *not* include? In various passages, Rawls takes special note of two categories of things that can be said to be just or unjust, yet are not the subject of his theory. One of these consists of the kinds of rules that regulate interactions and

transactions among private persons, such as those which regulate contractual agreements and those which apply to the practices of private associations (8/7). The other category is made up of individual actions and transactions. These things can certainly be said to be just or unjust, but they are not the subject of Rawls's theory. His topic is social justice, and in his view the appropriate subject of a theory of social justice is a society's basic structure.

Why focus on the basic structure of society? Rawls's main argument is that the institutions and practices that comprise a society's basic structure determine how well the members of that society are able to do in life, both in absolute terms and in comparison with others. In fact, in the most precise sense, it is the division of advantages that results from a society's basic structure rather than the basic structure itself that is the real subject of the theory (7/6). In identifying the basic structure as the primary subject of his theory of justice, Rawls was in effect adopting the view that justice is an attribute first and foremost of the terrain of society. For Rawls, the idea of justice applies principally to the landscape that determines the loci of privilege and deprivation in a society rather than to the character of relations among persons.

We can glean some additional features of Rawls's argument for focusing on the basic structure if we look at the following passage:

> The basic structure is the primary subject of justice because its effects are so profound and present from the start [. . .] men born into different positions have different expectations of life [. . .] the institutions of society favor certain starting places over others. These [. . .] inequalities [. . .] affect men's initial chances in life; yet they cannot possibly be justified by an appeal to the notions of merit or desert. (7/7)

This passage reveals two significant points. First, when Rawls argued for the basic structure as the appropriate subject of a theory of social justice, it is evident that his concerns about inequalities were concentrated on inequalities in people's life chances – on the (differential) opportunities available to people – and not on ultimate outcomes. He

writes here of the different positions men are "born into," of their "starting places" and "initial chances." Second, the passage hints at the fact (made clearer in later discussions) that Rawls was concerned about the ways in which major social institutions shape individuals' aspirations and expectations, as well as about the ways in which those institutions determine the division of advantages. Even if they have similar objective opportunities, some people do less well than others in life because they have lower aspirations or expectations. These aspirations and expectations are themselves shaped by the basic structure of society, and these subjective disparities among people were as worrisome to Rawls as objective differences in opportunities.

Rawls's argument for focusing on the basic structure also alludes to the inadequacy of the notions of merit and desert. Although his primary target of criticism is classical utilitarianism, he also takes aim at the idea that goods should be distributed in accordance with moral desert (310–315/273–277). It would take us too far afield to explore the intricacies of his discussion of this point, but it is worth noting here that Rawls dismisses desert as something fundamental to social justice in much the same way as he once dismissed attempts to justify the practice of punishment on retributivist grounds. He essentially replaces the concept of desert with that of legitimate expectations, a concept that separates the goods to which the members of a society are entitled from the contributions they make to that society in roughly the same way in which the principle "from each according to his ability, to each according to his needs" severs any connection between contributions and benefits (310–311/273–274).

For Rawls, the basic structure is not merely one among several possible subjects of a theory of justice, and social justice is not merely one among several possible types of justice. Social justice is instead justice in the most comprehensive and fundamental sense. Rawls envisages a division of labor between the principles of justice that apply to the basic structure and the rules or criteria of justice that apply to all other subjects. This division of intellectual labor is similar to the division he once conceived between a utilitarian justification of

the practice of punishment and a retributivist set of rules designed to constitute that practice. The principles of social justice are distinct from the rules and criteria that apply to other subjects. This is why he says that the "way in which we think about fairness in everyday life ill prepares us for the great shift in perspective required for considering the justice of the basic structure itself." At the same time, those principles are also intellectually prior to these other rules and criteria and serve as a foundation for defensible ideas about justice with regard to other subjects. As he observes in *A Theory of Justice*, once we have a sound theory of social justice, "the remaining problems of justice [including those which have to do with transactions, with criminal actions and punishments, and with compensatory justice, among other subjects] will prove more tractable in the light of it" (8/7).

The distinction Rawls draws between the principles of justice that apply to the basic structure and the rules and criteria of justice that apply to other subjects serves an important substantive purpose for his theory of justice as a whole. Recall that one of Rawls's principal objections to utilitarianism is that it is based on a monistic conception of the good – in other words, that it fails to accord due recognition to the fact that human beings legitimately hold a plurality of conceptions of the good. In his view, classical utilitarianism is a "comprehensive" theory, that is, a moral theory that offers prescriptions for the design of human institutions as well as for the decisions individuals should make, and indeed for all subjects to which any moral theory can be applied. The strong distinction he draws between principles of justice that apply to the basic structure – in effect, to the terrain of the social world itself – and criteria of justice for other subjects enables him to leave room for a plurality of moral views about those other subjects, which he believes should be accommodated by a theory of social justice.

Rawls characterizes his theory of justice as an "ideal" theory. By an ideal theory of social justice he means a theory that depicts a perfectly just society (8–9/7–8). Another phrase he uses for ideal theory is "strict compliance theory," which he contrasts with "partial

compliance theory." Rawls does not intend to diminish partial compliance theory, which deals with such topics as punishment; justice in the initiation, conduct, and aftermath of war; the justification of civil disobedience, militant resistance, and revolution; and compensation for wrongdoing, among many others. These matters are, he observes, pressing and urgent. Rawls's claim is that only by understanding the characteristics of a perfectly just society (he typically uses the phrase "well-ordered society," although for him that phrase has a broader meaning, encompassing societies that are not perfectly just) can we obtain a systematic grasp of the basis on which we should approach questions about justice in the real world. He regards ideal theory as more fundamental than non-ideal theory because he believes that we can best devise solutions to problems of justice that arise in the non-ideal world if we have first developed a sound conception of the principles of justice that would apply under ideal circumstances.

III

Rawls begins to lay out the most basic ideas of his theory with the following words:

> Let us assume [. . .] that a society is a more or less self-sufficient association of persons who [. . .] recognize certain rules of conduct as binding [. . .]. Suppose further that these rules specify a system of cooperation designed to advance the good of those taking part in it. Then [. . .] a society [. . .] is typically marked by a conflict as well as by an identity of interests. There is an identity of interests since social cooperation makes possible a better life for all than any would have if each were to live solely by his own efforts. There is a conflict of interests since [. . .] they each prefer a larger to a lesser share. A set of principles is required for choosing among the various social arrangements which determine this division of advantages [. . .]. These principles are the principles of social justice. (4/4)

With this passage as a touchstone, let's now look briefly at the theory's central ideas.

The most rudimentary of all the ideas underlying Rawls's theory is the idea of society as a fair system of social cooperation among free and equal persons over time, from one generation to the next. He sometimes calls this the "most fundamental intuitive idea" of the theory. Rawls offers no argument to defend this idea. Instead, he assumes that his readers will accept it as a plausible and appealing point of departure and concentrates his creative energies on the construction of an argument on the basis of this idea rather than on its defense.

This idea, then, plays a role in his theory of justice as fairness that is similar to the role played by the fundamental intuitive ideas of geometry in geometric reasoning. Although he did not believe it possible to construct a robust and persuasive theory of justice through pure deduction, Rawls aspired to make the argument of his theory as much like moral geometry as possible. The fundamental ideas on which theories of this kind are based are neither true nor false, and it makes little sense to attempt to prove or disprove them. Ultimately, those ideas stand or fall because of their usefulness or lack thereof. If the propositions and theories that are based on those ideas yield plausible or compelling accounts of the subjects to which they are addressed, then the usefulness of those ideas has been demonstrated. If not, then the ideas in question may be discarded in favor of alternatives.

Rawls believed that the idea of society as a fair system of social cooperation would be appealing to his readers. For most of his career (into the early 1980s) he appeared to believe that this appeal would be universal, at least to readers who had grappled sufficiently with the arguments of his theory to grasp its main points correctly. In his later years he seemed to retreat from this assumption by suggesting that his theory is designed to appeal distinctively to people who inhabit cultures that have been shaped by democratic and liberal ideals.

It is worth noting in any case that there is nothing bland or anodyne about the proposition that society should be conceived as a fair system of social cooperation among free and equal persons. Rawls's theory is

built on a proposition that is in fact highly controversial, both in an historical and in a geographical sense. Aristotle, for one, would have been aghast at this claim. Insofar as he conceived of persons as bearers of worth, he believed that they are of radically unequal worth because they are categorically unequal in capabilities, so that the notion that we should think of society as a system of cooperation among equal persons would have made no sense to him. Nor would he have had much sympathy or appreciation for the emphasis this proposition places on freedom. For him, human beings are endowed with functions that are prescribed by nature. Excellence is exhibited through outstanding performance of those prescribed functions, much as excellence in acting is displayed through outstanding performance in a scripted role. Many pre-modern thinkers would have found the fundamental intuitive idea of society as a fair system of social cooperation among free and equal persons incomprehensible, and some would have found it reprehensible. The same can be said of many people today who have escaped the influence of, or rejected, modern European ideas (it can also be said of some people who embrace modern anti-liberal European ideas). On an historical and worldwide scale, the foundation on which Rawls constructed his theory is by itself a radical proposition.

For Rawls, the idea of society as a fair system of social cooperation is a basis for reasoning about societies in what he, following David Hume, calls the circumstances of justice (126–130/109–112). The circumstances of justice are circumstances of moderate scarcity, in which the hand of nature is neither so generous as to give human beings all they want, with no need for labor or social cooperation, nor so harsh as to force people into a struggle for survival so elemental as to preclude social cooperation. The circumstances of justice are those in which we neither enjoy unlimited abundance nor suffer extreme deprivation.

If the fundamental idea of Rawls's theory is that of society as a fair system of social cooperation among free and equal persons, the key question of that theory is: on what terms should this cooperation

proceed? For the purposes of his theory of social justice, Rawls thinks of society as a collaborative enterprise of a sort that is akin to a business partnership, a "cooperative venture for mutual advantage." (He did not, however, think of society as a voluntary association, because for the most part membership in societies is thrust upon individuals who have little or no chance either to grant or to withhold their consent.) This conception of society is rooted in Adam Smith's contention that a complex division of labor is the principal source of the great wealth of modern societies. For Rawls, questions about social justice arise as a result of the productivity, broadly construed, that is made possible by the division of labor. As he says, "social cooperation makes possible a better life for all than any would have if each were to live solely by his own efforts." Society is a sort of partnership that is undertaken for the mutual benefit of those who enter into – or in this case, typically find that they are already in – that partnership. The key question of social justice is a question about the terms of this partnership, and in particular about the way in which its benefits will be distributed among the participants.

From this conception it follows that, for Rawls, the distributive questions to which the idea of social justice points focus distinctively on the social product, that is, on the "goods" (in a broad sense) that are generated by the joint efforts of the partners. These goods may not all be "material" or "economic" goods of the sort Smith had in mind. For example, they may include enjoyments of a non-economic kind that can be achieved only through collaboration with others, such as the enjoyments we derive from participating in a game that requires a number of participants, or from friendship. It is for these goods – the diverse class of goods that are generated by the joint efforts of the partners – and for these goods alone, however, that we require a set of principles to determine the proper distributive shares.

Rawls's key question is a variation of the question of social justice Sidgwick had raised roughly a century earlier, namely whether any clear principles may be found on the basis of which we can discover an

ideally just distribution of rights, privileges, burdens, and pains. Notice, however, that, whereas Sidgwick had raised this question about a distribution of these things "among human beings as such," Rawls narrows the question to one about the distribution of advantages among the members of a given society conceived as a cooperative venture for mutual advantage. Rawls appears to have believed that we can find a compelling set of principles of social justice only by restricting the scope of our inquiry to a particular (even if hypothetical) society rather than by extending it to all humankind.

Notice also that, while Sidgwick had written with equal emphasis about the distribution of burdens and pains as well as that of rights and privileges, Rawls's emphasis is decidedly on the division of advantages. One reason for this emphasis may lie in his conception of society as a mutually advantageous undertaking. While some members benefit far more than others, Rawls supposes that normally all are made better off by participating in a scheme of social cooperation than they would be if "each were to live solely by his own efforts." So the significant thing that is generated by a scheme of social cooperation is benefits, not burdens; and it is the things generated by social cooperation that are subject to principles of social justice.

A second and more interesting reason, however, may lie in his assumption that all the members of such an enterprise will be active participants not only in the narrow sense of adhering to its rules of conduct, but also in the wider sense of contributing to it by being normal cooperating members. The principal conceptions of social justice that were developed during the long nineteenth century either offered a prescription for the contributions members should make to society and for the benefits they should receive (*from* each according to his ability/*to* each according to his needs) or linked the benefits individuals should receive to their contributions (the principle of desert). In contrast, Rawls's theory focuses on benefits while bracketing questions about contributions. Rawls appears simply to assume that the members of a just society will contribute to that society's social product in accordance with their diverse talents. This assumption

seems to be part of what he intends when he suggests that the members of a society that is based on a fair system of cooperation among free and equal persons over time, from one generation to the next, would be "normal cooperating members of society over a complete life."

To find an answer to his central question, Rawls adopts a method that is borrowed in part from Kant and some of his predecessors in early modern political thought, including Thomas Hobbes, John Locke, and Jean-Jacques Rousseau. The method is to imagine that a society has been founded by an agreement among its members that determines the terms of their association. Kant had employed this method in his theory of public right by invoking the idea of the original contract to test the justice of public laws and policies. If it is plausible to suppose that the law or policy in question would have received the approval of all the members of a society in an original contract, then according to Kant we must assume that that law or policy is just. If this supposition is implausible, then we may conclude that the law or policy is unjust.

Kant limited his use of the idea of a hypothetical original contract to the task of testing the justice or injustice of laws and policies. In contrast, Rawls uses the idea of a hypothetical contract to identify a set of principles of social justice. Rawls's use of this device is more ambitious and more elaborate than Kant's.

Rawls asks his readers to imagine that each member of society is represented by an agent in a condition he calls the "original position," a hypothetical state of affairs in which the agents come together to reach an agreement that will shape the terms on which the society operates. The object of the agents' agreement (adopting legalistic language, Rawls typically calls these agents the "parties" in the original position) will be a set of principles of social justice focused on the distribution of advantages in society. Once these principles have been adopted, they can be used in a second stage of deliberation, which he called the stage of the constitutional convention, to make a choice among the various alternative basic structures that are available to the society. The basic structure they select will in turn provide the

framework within which laws will be adopted, policies developed, and specific decisions reached. Since his entire theory of social justice is an ideal theory, these principles of justice will of course be framed for a perfectly just society.

Because he wants his readers to imagine a hypothetical contract that will be far more ambitious (in the sense of doing more intellectual work) than Kant's idea of the original contract, Rawls provides a significantly more detailed description of the original position than Kant does of the original contract. He emphasizes that the parties in the original position are rational in the sense that they prefer for the members of society whom they represent to obtain a greater rather than a lesser share of the benefits of social cooperation. The fact that the parties are rational does not entail that they or the members of society whom they represent are egoistic. Those members may, for example, wish to use a portion of their shares to promote causes that benefit others. He also emphasizes that the parties are reasonable. They understand that they must be willing to reach agreement with their counterparts on fair terms. In order to help guarantee their reasonableness, Rawls asks us to imagine that the parties in the original position have been placed behind (what he calls) a "veil of ignorance" that prevents them from knowing the abilities, social positions, or indeed the very identities of the members they represent. This kind of knowledge might sway them to bargain for unfair advantages. For example, if a representative were to know that the member he represents is intellectually exceptional, he or she might demand principles of justice that would tend to favor the intellectually gifted. Finally, Rawls suggests that the parties in the original position would adopt a distinctive measure to determine how well-off the members they represent are in comparison with others. The measures that are most commonly used for this purpose are income and wealth. Classical utilitarians used happiness (though they usually supposed that people with greater income or wealth are happier than others). Rawls argues that the appropriate measure would be made up of several diverse elements, including certain rights and liberties, income

and wealth, and the social bases of self-respect, elements that he called "social primary goods."

Rawls's proposal, then, is that we can discover the best set of principles of social justice by imagining a number of representatives, in the hypothetical scenario he calls the original position, who want to reach an agreement with one another that will best serve their clients' interests, where their "clients" are the members of the perfectly just society that will be brought into being on the basis of those principles. As we have seen, it is a premise of Rawls's theory that some members of such a society – not merely of any actual society, but of a perfectly just society – will be better off than others. And not only that: some will be born into different positions, develop different expectations, and be endowed with different chances in life from others. Just as he borrowed from Adam Smith the idea that the division of labor is by far the most important source of productivity and ultimately of wealth, so did he also inherit from some of the classical political economists the assumption that the same division of labor leads ineluctably to disparities in the opportunities available to the different members of a society. Rawls assumed that human beings are equal to one another in worth. That assumption is one of the points conveyed by his beginning with the fundamental intuitive idea of society as a fair system of social cooperation among free and equal persons. But he also assumed that all members of a society can benefit from advantages obtainable only through a complex division of labor and that inequalities are an inevitable by-product of such a division of labor. Rawls's premises are egalitarian, but the principles of social justice at which he arrives are designed to justify those inequalities which (he believed) work to the advantage of all.

IV

The principal conclusion of the theory of justice as fairness is that the terms of social cooperation that would constitute the basic principles

of social justice in a perfectly just society – the terms to which the parties in the original position would agree – can be summarized as follows:

1 Each person has an equal right to a fully adequate scheme of equal basic liberties which is compatible with a similar scheme of liberties for all.
2 Social and economic inequalities are to satisfy two conditions. First, they must be attached to offices and positions open to all under conditions of fair equality of opportunity; and second, they must be to the greatest benefit of the least advantaged members of society.

The first of these principles has (what Rawls calls) lexical priority over the second, and the first part of the second principle has lexical priority over its second part. In other words, the first principle must be fully satisfied before the second comes into play, just as all the words that begin with the letter "a" are listed in a dictionary before the words that begin with the letter "b." The way in which a society's social and economic inequalities are distributed is relevant for an evaluation of the justice of that society's basic structure only when all its members enjoy a fully adequate scheme of liberties. Let us call the first of these requirements the *basic liberties principle*. Since the second principle has two parts, let's call its first part the *equal opportunity principle* and the second (following Rawls's own consistent usage) the *difference principle*.

Rawls also reaches a second major conclusion, to which he believes the parties in the original position would agree. He argues that, in addition to a set of principles that can be invoked to choose among alternative basic structures (the two principles of justice specified just above), those parties would agree that the members of a just society should possess certain attributes. First, they would want the members (citizens) of such a society to possess an effective sense of justice. By this he means that they would want those members to be able to

understand, to apply, and to act on the basis of a set of public principles of justice (namely the two principles of justice as fairness). Integral to his conceptions of social justice and of a well-ordered society is the conviction that, for a society truly to be just, its members must understand and consent to the terms of social cooperation by which they are governed. This first point underscores this conviction. Second, they would want the members to possess and to develop their capacities for a conception of the good. In other words, each member of society would want all the others to develop the capacity to form, to revise, and rationally to pursue a conception of the good, a conception that would form the basis of a member's plan of life. Rawls calls these attributes the two "highest-order" moral powers because they are the attributes he believes the parties in the original position would most want the citizens of a perfectly just society to develop. He labels the theory that describes these powers "the theory of moral personality." So for Rawls the theory of justice as fairness specifies both a set of principles of social justice (the two principles of justice as fairness) to which we should turn in choosing among alternative basic structures of society, and a set of attributes (the attributes of moral personality) that a just society should cultivate in its members.

These two major conclusions are intertwined. For example, the liberties that are to be protected by the basic liberties principle are, according to Rawls, just those liberties which are essential to the development and exercise of the two highest-order powers of moral persons. Rawls does not attempt to supply a complete list of these liberties, but he mentions, among others, freedom of thought, liberty of conscience, and freedom of association; personal liberties such as freedom from arbitrary arrest, the rights to due process of law and to a fair trial; and political liberties such as the right to vote and freedom of the press. He lays special emphasis on political liberties, and insists that the members of a just society must enjoy the "fair value" of these liberties, by which he means that each member should be in a position to exercise as much influence over common decisions as any other member.

The equal opportunity principle entails that the positions or roles in society to which unequal rewards are attached must be open to fair competition on a basis of equal opportunity. This principle should be underwritten by education for all, among other measures.

The difference principle prescribes that social and economic inequalities are justified only insofar as they work to the benefit of the least advantaged members of society. At first glance, the notion that such inequalities might be beneficial to the least advantaged – to the members of society who enjoy the most restricted opportunities and the fewest resources – seems paradoxical. Remember, however, that Rawls inherited from earlier political economists the assumption that the same division of labor that accounts for the unprecedented productivity and wealth of modern societies also leads inevitably to disparities in the opportunities available to different members of society. If the increase in goods (wealth and other goods, as measured by an index of social primary goods) made possible by a complex division of labor is sufficiently great, then even the least advantaged members of a society might be better off in a basic structure in which that complex division of labor prevails than they would be in an alternative basic structure without it. The difference principle takes this possibility into account.

The difference principle is the most distinctive of all the conclusions of the theory of justice as fairness. Here as elsewhere, Rawls's principal target is the theory he considers the most serious rival to the theory of justice as fairness, namely classical utilitarian theory. To see why, consider an illustration. Suppose you are a member of a society of one hundred members. Assume that the well-being of each of those members, as measured by an index of social primary goods, can be expressed on a cardinal scale of 1 to 10, with 10 representing the highest possible standard of well-being (as measured by one's share of primary goods) and 1 representing the lowest possible share. (In a cardinal scale, a share of 4 has twice the value of a share of 2 and a share of 8 has twice the value of a share of 4, while a share of 8 is more valuable than one of 7 by just the same amount as a share of 5 is more

valuable than a share of 4. Having a single share is like having a single orange, or a single unit of any other good, while having three shares is like having three oranges.) Now imagine that your society is faced with a choice between two alternative basic structures whose distributive consequences can be represented as follows:

Shares of Primary Goods	Basic Structure A	Basic Structure B
10		
9	25 persons	
8		
7		25 persons
6	50 persons	
5		50 persons
4		25 persons
3	25 persons	
2		
1		

In Basic Structure A, then, 25 of the society's 100 members enjoy 9 shares of primary goods each, while 50 enjoy 6 shares each, and 25 must make do with 3 shares each.

If we think of shares of primary goods as units of well-being, then it is easy to see that the aggregate well-being that would be enjoyed by the members if they were to adopt Basic Structure A can be represented by the figure 600 $[(25 \times 9) + (50 \times 6) + 25 \times 3)]$, while a similar calculation will show that Basic Structure B would yield an aggregate well-being of 525. If we suppose, for the sake of the argument, that well-being as measured by shares of primary goods is equivalent to well-being as measured by utility, then it is clear that the greatest happiness principle would direct us to adopt Basic Structure A. Yet the difference principle would prescribe the adoption of Basic Structure B,

since under that structure the least advantaged members of society are better off (at 4 on the scale of primary goods) than they would be under Basic Structure A (which would leave them at 3 on the scale). Basic Structure B leads to inegalitarian consequences, but those consequences are more advantageous to the least advantaged members of society than the consequences of the alternative.

This illustration assumes that Basic Structures A and B are the only available alternatives. If any additional option were available that would leave the least advantaged members of society even better off than they would be under Basic Structure B, then the difference principle would prescribe that, as a matter of social justice, we should adopt that third basic structure. For example, if the range of possible basic structures included a Basic Structure C under which all the members of society would command identical shares of primary goods rated at 5 on our scale of 1 to 10, then the difference principle would direct us to adopt that structure, even though aggregate well-being under it would be lower than under either of the alternatives [$100 \times 5 = 500$], because the least advantaged members of society would be better off under Basic Structure C than under either A or B. Because of his assumptions about productivity and the division of labor, Rawls did not seem to believe that such an alternative would be possible, but the principles of justice as fairness do not rule it out.

Rawls's standard statement of the difference principle seems slightly discrepant with his defense of the basic structure as the appropriate subject of the theory of social justice. The difference principle, which states that "social and economic inequalities [. . .] must be to the greatest benefit of the least advantaged members of society," suggests a focus on ultimate outcomes, that is, on how well off (as measured by shares of primary goods) the members of a society turn out to be. Yet Rawls's defense of the basic structure as the primary subject of his theory of justice focuses on opportunities ("starting places"), not on ultimate outcomes. In fuller statements of the difference principle, he sometimes speaks of the "greatest *expected* benefit"; and it is evident,

in various places in his work, that Rawls understood the important difference between initial chances and ultimate outcomes. In his discussions of the principles of justice as fairness, however, Rawls frequently elides this distinction.

V

The theory of justice as fairness is an extraordinary accomplishment. As a vision of social justice for a society whose members are presumed to be free and equal citizens, it has no peer. Nevertheless, the theory is not flawless. I shall focus my comments on the way in which Rawls construes the subject of his theory.

Rawls's assertion that the basic structure of society is the appropriate subject of a theory of social justice is widely understood to be one of the most distinctive claims of his theory. As we have seen, the claim is not merely that the basic structure happens to be the appropriate subject of a theory of social justice in the same way in which (say) law violations are the appropriate subject of a theory of penal justice. It is rather that the basic structure has a kind of priority over all other kinds of subjects pertaining to justice, so that social justice is justice in the most comprehensive and fundamental sense. For Rawls, a sound theory of social justice provides the necessary foundation on which we can construct solutions to other, less comprehensive problems of justice. (In the latter years of his career, Rawls took up a set of questions about justice beyond national borders, questions that are arguably as comprehensive as, or more so than, questions about social justice within borders.)

If we examine Rawls's arguments closely, we can see that his claim consists of three distinct parts. The first is a causal claim that the institutions and practices that comprise a society's basic structure determine how well the members of a society are able to do in life. The second is the conceptual claim that the principles of justice that apply

to the basic structure may be quite different in character from the rules and criteria that apply to other problems of justice. The third is a claim of intellectual priority. The claim is that we can best address the wide range of questions that arise about justice by first developing a sound theory of social justice. This theory can then constitute the foundation for defensible ideas about justice with regard to other subjects.

The first of these claims, in a general form, is incontrovertible. How completely a society's basic structure determines how well its members are able to do may be controversial, but there can be little doubt that a society's major institutions have profound effects on its members and on the division of advantages among them.

It is not difficult to see the force of Rawls's second claim as well. Consider the example of labor contracts. In a society made up of employers who are small business owners with limited resources and employees who are independent proprietors with a significant range of employment opportunities from which to choose, we can expect that justice will be served if all parties are free to enter into labor contracts on whatever terms are mutually agreeable. Since all parties possess roughly equal bargaining power, the bargains they reach typically can be expected to be fair. Matters will be different in a society dominated by giant corporate employers with vast resources at their command and by employees who have few alternatives (or, in the limiting case of some company towns, only one serious employment opportunity). Because of the great disparities in bargaining power in the latter scenario, freedom of contract is likely to lead to labor agreements that are unfair to employees. In that case collective bargaining arrangements, which reduce disparities in bargaining power between employees and employers, may restore some balance and justice to the labor contracts to which the parties agree. (In some cases, of course, collective bargaining arrangements may confer excessive power on those who bargain on behalf of employees.) A significant shift in perspective is required to grasp the fact that, in situations of great disparity in bargaining power, fairness is best secured by arrangements that differ sharply from those which typically

lead to fair bargains in situations of relatively equal bargaining power. It is not surprising that a similar or greater shift in perspective may be required to grasp the fact that fair principles of justice for the basic structure of a society may differ markedly from the rules or criteria of justice that apply to ordinary interactions among individuals.

The claim that the principles of social justice are intellectually prior to, and serve as a foundation for, defensible ideas about justice with regard to other subjects is more problematic. Consider for another brief moment the example of labor contracts. If agreements reached by employers and employees who possess roughly equal bargaining power under conditions of freedom of contract are likely to be fair, the reason for this fact is that those agreements will typically embody the norm of balanced reciprocity. If collective bargaining arrangements help to restore fairness under conditions of highly unequal bargaining power, the reason is that those arrangements bring labor agreements more nearly into line with the norm of balanced reciprocity.

Nothing is more central to the way in which human beings think about fairness among relative equals than the norm of balanced reciprocity. In a chapter in *A Theory of Justice* on "The sense of justice," Rawls observes:

> reciprocity, a tendency to answer in kind [. . .] is a deep psychological fact [. . .]. A capacity for a sense of justice built up by responses in kind would appear to be a condition of human sociability. (494–495/433)

The kind of reciprocity Rawls has in mind here is balanced reciprocity, "a tendency to answer in kind." Although the justice of collective bargaining arrangements is not intuitively obvious to most people, the argument for the justice of those arrangements rests on intuitions that are highly accessible as well as widely, perhaps even universally, shared. The same thing can be said of the principles of social justice, as Rawls seems to acknowledge when he observes that the "most stable conceptions of justice are presumably those for which the corresponding sense of justice is most firmly based on these tendencies" (495/433).

In short, while it seems sensible to claim both that a society's basic structure plays a large causal role in determining how well its members are able to do and that the principles of social justice may be distinct from those which apply to other subjects, it is misleading to suppose that the principles of social justice are intellectually prior to and constitute the foundation for ideas about justice in relation to all other subjects. The kind of justice that applies directly to relations among persons is not trumped by the principles of social justice. Instead, the principles of social justice are rooted in the idea of justice in direct relations among persons. This idea – that justice among relative equals is based on the norm of balanced reciprocity – possesses an integrity that is not overshadowed by, and in fact provides the intellectual foundation for, sound ideas about social justice. Principles of social justice are distinct from the principles that apply to direct relations among relative equals, because the complexity of social institutions and practices requires adjustments to those principles. Ultimately, however, sound principles of social justice will be based on the norm of balanced reciprocity among relative equals.

If sound ideas about social justice are rooted in the norm of balanced reciprocity, then the concept of desert, which Rawls dismisses perfunctorily, may have a role to play in the way we think about justice, including social justice, after all. If two persons, A and B, are relative equals, and A confers a benefit on B, then there is a sense in which A deserves to be requited with a benefit similar in value to the benefit she has conferred, and B has an obligation of justice to bestow a benefit on A in return for the benefit he has received. Similarly, if Q inflicts a harm on R, then there is a sense, independent of any particular conception of social justice, in which Q deserves to suffer some harm in return.

Of course, the norm of balanced reciprocity in its simplest form – the form that applies to bilateral relations between relative equals – is not adequate as a guide to justice in relations among persons in complex circumstances. In situations that are multilateral or in which people are unequally placed, the social arrangements that would lead

to justice in relations among persons may be dramatically different from those which apply to simple bilateral relations between equals. To accommodate these situations, major adjustments are needed, in much the same way as adjustments are required in bargaining between employees and employers when the disparities in bargaining power between them are large.

We can therefore see how the *concept* of desert might play a significant role in the way we think about justice, without leading us to endorse either the *principle* of desert (the contribution principle) or retributivist reasoning in its classic form (the form that is based on strict balanced reciprocity between putative equals). Rawls was right to see that the principles of justice that apply to the basic structure of a society are conceptually distinct from the rules of justice that apply to simple bilateral relations between persons. In fact his insight is generalizable to many subjects in addition to the basic structure of society. Yet, regardless of the particular subject for which the principles of justice are designed, if they are to be recognizable and acceptable to human beings, they must be rooted in the sense of justice – a sense that is best expressed through the concepts of reciprocity and desert.