CHAPTER 25

A just social contract

> A republican constitution is a constitution which is founded upon three principles. First, the principle of the *freedom* of all members of a society as men. Second, the principle of the *dependence* of all upon a single common legislation as subjects, and third, the principle of the *equality* of all as *citizens*. This is the only constitution which is derived from the idea of an ongoing contract upon which all rightful legislation of a nation must be based. (Italics in original)
>
> Immanuel Kant

One of the most influential studies of the first stages of the social choice process has been Rawls's *A Theory of Justice* (1971). This book is at once a contribution to moral and to political philosophy. Rawls relies on work and results appearing in various branches of the social sciences, however, and applies his theory to several of the major issues of the day. For this reason, Rawls's work has been widely read and discussed and has had a substantial impact on the economics literature in general, and on collective choice in particular.

Rawls's theory differs from those that we have discussed up to now in its focus on the *process* or *context* in which decisions are made as much as, if not more than, on the outcomes of this process. The goal is to establish a set of just institutions in which collective decision making can take place. No presumption is made that these institutions or the decisions emerging from them will in any sense maximize the social good (pp. 30–1, 586–7).[1] Here we see a clear break with the social welfare function approach. More generally, Rawls challenges the utilitarian philosophy that underlies the SWF methodology and that has reigned in discussions of these topics over the past two centuries.[2]

Rawls sets out to develop a set of principles to apply to the development of "the basic structure of society. They are to govern the assignment of rights and duties and regulate the distribution of social and economic advantages" (p. 61). These principles form the foundation of the social contract, and Rawls's theory is clearly

[1] This and all subsequent page references in this chapter are to Rawls (1971) unless otherwise indicated.

[2] Bruce Ackerman (1980) is critical of both utilitarianism and contractarianism as approaches to deriving principles of justice. Instead, he emphasizes dialogue as the *process* by which these principles are established.

His criticism of contract theory seems overdrawn, however. Unless dialogue eventually leads to a consensus on the principles that underlie the liberal state, the liberal state can never come into being. If agreement on principles is ultimately achieved, that agreement becomes a form of social contract that binds the citizens of the liberal state together. Dialogue is an important part of the process by which agreement is obtained, but not a substitute for the agreement.

one of the major, modern reconstructions of the contractarian argument. The theory is developed in two parts: first, the arguments in favor of the contractarian approach are established. Here the focus is upon the characteristics of the original position from which the contract is drawn. The moral underpinning of the social contract rests on the nature of the decision process taking place within the *original position*, which in turn depends upon the setting in which the original position is cast. The second part of the theoretical argument develops the actual principles embedded in the social contract. Rawls emphasizes the independence of these two arguments. One can accept either part without necessarily committing oneself to the other (pp. 15 ff.). This point is important to keep in mind since the different parts have been attacked in different ways and one might feel more comfortable about one set of arguments than another. This two-part breakdown forms a natural format by which to review Rawls' theory. Following this review, we examine some of the criticisms of the theory that have been made.

25.1 The social contract

Perhaps the easiest way to envisage how the social contract comes about in Rawls's theory is to think of a group of individuals sitting down to draw up a set of rules for a game of chance, say, a game of cards, in which they will subsequently participate.[3] Prior to the start of the game, each individual is ignorant of the cards to be dealt to him and uncertain of his skills relative to those of other players. Thus, each is likely to favor rules that are neutral or fair with respect to the chances of each player, and all might be expected to agree to a single set of fair rules for the game. Here the incentive "to get on with the game" can be expected to encourage this unanimous agreement.

In Rawls's theory, life is a game of chance in which Nature deals out attributes and social positions in a random or accidental way (pp. 15, 72, 102 ff.). Now this natural distribution of attributes and chance determination of social position is neither just nor unjust (p. 102). But it is unjust for society simply to accept these random outcomes, or to adopt institutions that perpetuate and exaggerate them (pp. 102–3). Thus, a set of just institutions is one that mitigates the effects of chance on the positions of individuals in the social structure.

To establish such a set of institutions, individuals must divorce themselves from knowledge of their own personal attributes and social positions by stepping through a *veil of ignorance* that screens out any facts that might allow an individual to predict his position and benefits under a given set of principles (pp. 136 ff.). Having passed through the veil of ignorance, all individuals are in an *original position* of total equality in that each possesses the same information about the likely effects of different institutions on his own future position. The original position establishes a status quo of universal equality from which the social contract is written (pp. 3–10).

Individuals in the original position about to choose a set of principles to form a social contract resemble individuals about to draw up rules for a game of

[3] The analogy between a social contract or constitution and drawing up rules for a parlor game is often used by Buchanan. See, for example, Buchanan (1966) and Buchanan and Tullock (1962, pp. 79–80).

chance – with one important difference. Individuals choosing rules for a game of chance are ignorant of their future positions by necessity, and thus can be expected to adopt fair rules out of self-interest. Individuals in the original position are ignorant of their present and likely future positions, because they consciously suppress this information by voluntarily passing through the veil of ignorance. Although they may choose institutions out of self-interest once they are in the original position, the act of entering *the original position* is a moral one, whose ethical content rests on the argument that information about the distribution of certain "factors [is] arbitrary from a moral point of view" (p. 72). Justice is introduced into the social contract via the impartiality incorporated into the collective decision process through the nature of the information made available to individuals in the original position. Thus emerges the fundamental notion of *justice as fairness*.

What, then, is the nature of the information screened out by the veil of ignorance? Rawls's views here are rather strict. Not only is knowledge of their natural talents, tastes, social position, income, and wealth denied them, but also information about the generation to which they belong, the state of economic and political development of their society, and other fairly general information that Rawls argues might nevertheless bias an individual's choice in the direction of one set of principles over another. For example, knowledge of the generation in which an individual lives might lead him to favor a particular type of public investment policy, or social discount rate, thereby benefitting his generation at the expense of others. Given the very general nature of the information that individuals have in the original position, it is plausible to assume that the principles on which they agree are impartial with respect to the advantages they provide, not only for specific individuals, or individuals in well-defined positions, but even for individuals in different generations and living under different economic and political systems. Since all individuals have access to the same information once they have passed through the veil of ignorance, all will reach the same conclusions as to the set of just principles that ought to be embedded in the social contract. Equality in the original position leads to unanimity over the social contract.

25.2 The two principles of justice

Given the information available in the original position, Rawls argues that the following two principles will be chosen as the pillars of the just social contract:

> *First:* each person is to have an equal right to the most extensive basic liberty compatible with a similar liberty for others.
> *Second:* social and economic inequalities are to be arranged so that they are both (a) reasonably expected to be to everyone's advantage, and (b) attached to positions and office open to all. (p. 60) [These] two principles (and this holds for all formulations) are a special case of a more general conception of justice that can be expressed as follows. All social values – liberty and opportunity, income and wealth, and the bases of self-respect – are to be distributed equally unless an unequal distribution of any, or all, of these values is to everyone's advantage. (p. 62)

It is perhaps intuitively obvious that something like the "more general conception of justice" appearing on page 62 would emerge from a collective decision process in which the individuals were ignorant of their future positions and thus were induced to act impartially. Indeed, in some ways the setting of the original position resembles the familiar cake-cutting problem in which one individual divides the cake and the other chooses the first piece. By analogy with this example, one would expect the principles emerging from the original position to have an egalitarian tone, as is present in the more general conception. Rawls adds flesh to his theory, however, by deriving the two, more specific principles quoted above as part of the *special* conception of justice that is thought to hold once a society has reached a point of moderate scarcity, and by further arguing that these two principles will be chosen in lexicographical order. The first principle always has precedence over the second (pp. 61 ff., 151 ff., 247–8).

Rawls defends the lexicographical ordering of these two principles as follows:

> Now the basis for the priority of liberty is roughly as follows: as the conditions of civilization improve, the marginal significance for our good of further economic and social advantages diminishes relative to the interests of liberty, which become stronger as the conditions for the exercise of the equal freedoms are more fully realized. Beyond some point it becomes and then remains irrational from the standpoint of the original position to acknowledge a lesser liberty for the sake of greater material means and amenities of office. Let us note why this should be so. First of all, as the general level of well-being rises (as indicated by the index of primary goods the less favored can expect) only the less urgent wants remain to be satisfied by further advances, at least insofar as men's wants are not largely created by institutions and social forms. At the same time the obstacles to the exercise of the equal liberties decline and a growing insistence upon the right to pursue our spiritual and cultural interests asserts itself. (pp. 542–3)

Thus, Rawls sees society as better able to "afford" the extension of equal liberties to all citizens as it develops; that is, he sees liberty as essentially a luxury good in each individual's preference function. With increasing levels of income, the priority of liberty over other psychological and material needs rises, until at some level of development it takes complete precedence over all other needs.

The second principle of justice, which Rawls names the difference principle, also contains a lexicographic ordering. The welfare of the worst-off individual is to be maximized before all others, and the only way inequalities can be justified is if they improve the welfare of this worst-off individual or group. By simple extension, given that the worst-off is in his best position, the welfare of the second worst-off will be maximized, and so on. The difference principle produces a lexicographical ordering of the welfare levels of individuals from lowest to highest. It is important to note that Rawls defines welfare levels not in terms of utility indexes or some similarly subjective concept, but in terms of *primary goods*. These are defined as the basic "rights and liberties, powers and opportunities, income and wealth" that a society has to distribute (p. 62; see also pp. 90–5). Here we have another

Table 25.1. *Payoff possibilities*

	W	B
S_1	0	n
S_2	$1/n$	1

example of the break that Rawls is trying to establish between his theory and classical utilitarianism. The principles embedded in the social contract must be general. They must apply to all and be understandable by all (p. 132). This requirement places a bound on the complexity that can be allowed to characterize the basic principles of the social contract. The lexicographical nature of the difference principle and its definition in terms of objectively discernible primary goods make it easy to apply.

The difference principle is closely related to the maximin strategy of decision theory. This strategy dictates that an individual should always choose the option with the highest minimum payoff regardless of what the other payoffs are or the probabilities of obtaining them. The force of the strategy can easily be seen in an example Rawls himself uses when discussing the principle (pp. 157–8). Let W and B be two possible states of the world, say, the drawing of a white or black ball from a sack. Let S_1 and S_2 be the strategy options with prizes as given in Table 25.1. The maximin strategy requires that one always pick strategy S_2, regardless of the value of n and regardless of the probability, p, of a white ball being drawn, as long as $n < \infty$, and $p > 0$. One will never pay an amount, however small, to win a prize, however large, no matter what the probability of winning is, as long as it is not a sure thing.

Given the conservatism inherent in the maximin decision rule, Rawls goes to great pains to rationalize incorporating this rule into his basic principle of distributive justice. His reasons are three:

> First, since the rule takes no account of the likelihoods of the possible circumstances, there must be some reason for sharply discounting estimates of these probabilities. (p. 154)
>
> Now, as I have suggested, the original position has been defined so that it is a situation in which the maximin applies [and] the veil of ignorance excludes all but the vaguest knowledge of likelihoods. The parties have no basis for determining the probable nature of their society, or their place in it. Thus they have strong reasons for being wary of probability calculations if any other course is open to them. They must also take account of the fact that their choice of principles should seem reasonable to others, in particular their descendants, whose rights will be deeply affected by it. (p. 155)
>
> The second feature that suggests the maximin rule is the following: the person choosing has a conception of the good such that he cares very little, if anything, for what he might gain above the minimum stipend that he can, in fact, be sure of by following the maximin rule. It is not worthwhile for him to take a chance for

> the sake of a further advantage, especially when it may turn out that he loses much that is important to him. This last provision brings in the third feature, namely, that the rejected alternatives have outcomes that one can hardly accept. The situation involves grave risks. (p. 154)

Thus Rawls's arguments for the difference principle rest heavily upon his assumptions about the information available in the original position, and the economic conditions facing society. Society is in a state of "moderate scarcity"; the poor can be made better off without great sacrifice to the rich (pp. 127–8). The assumption of moderate scarcity also plays an important role in justifying the lexicographic priority of the liberty principle over the difference principle, as already noted (pp. 247–8). Obviously, situations could be envisaged in which an individual would be willing to give up a certain degree of liberty for an increase in material goods, or risk being slightly poorer for a chance to be substantially richer. Rawls assumes, however, that the marginal utility of material gains declines rapidly enough as prosperity increases, and that society is already wealthy enough, so that these trade-offs and gambles at unknown odds are no longer appealing.

25.3 Extensions of the theory to other political stages

Rawls extends his theory to consider the characteristics of subsequent stages in the political process: the constitutional stage, the parliamentary stage, and administrative and judicial stages. In each subsequent stage, the veil of ignorance is lifted to some extent and individuals are given more information with which to make collective decisions. For example, in the constitutional stage, individuals are allowed to know the type of economic system with which they are dealing, the state of economic development, and so on. At each subsequent stage, however, knowledge of specific individual positions and preferences are denied to individuals making collective decisions. Impartiality is thus preserved, and the two principles of justice continue on into subsequent stages of the political process in precisely the same form in which they appear in the social contract. Thus, the social contract forms the ethical foundation for all subsequent political stages. As with the social contract stage itself, Rawls does not envisage actual political processes at work, but rather a form of *Gedanken-experiment* in which individuals reflect upon the principles that *ought* to underlie the social contract, constitution, or subsequent stages. In the original position, as defined for the constitutional stage, a hypothetical, just constitution is drafted in the same way that a hypothetical, just social contract is drafted by individuals at this earlier stage. This just constitution, once drafted or conceptualized, can then be compared with actual constitutions to determine in what respect they are in accord with the ethical principles contained in this hypothetical constitution. Of course, once one has specified the principles underlying a just constitution, and assuming that all can agree on them, one would be free to redraft actual constitutions to conform to these principles. But the leap from hypothetical constitutions formulated introspectively to actual constitutions written by individuals with real conflicts of interest may be a great one.

25.4 Critique of the Rawlsian social contract

A *Theory of Justice* has precipitated so much discussion and critical evaluation that we cannot hope to survey all of this material here. Instead, we focus on those issues that are most relevant to the public choice literature. Again the material can be most easily organized around Rawls's arguments in favor of the contractarian approach and the two principles underlying the contract formed.

25.4.1 *The social contract*

Until the appearance of Rawls's book, social contract theory had fallen into disrepute. The historical version of the theory had been fully discredited for over a century, and as a purely theoretical account for the existence of the state it was thought by many to be redundant.[4] This latter criticism is certainly valid from a public choice perspective. The theory of public goods, the prisoners' dilemma, externalities, the existence of insurable risks, and a variety of similar concepts suffice to explain why individuals might out of self-interest reach unanimous collective agreements. Now a contract is nothing more than a unanimous collective agreement to the provisions specified in it. Thus, any decision that can be explained via the creation of a contract can probably be explained just as well as a unanimous collective decision (vote). Not all public good and prisoners' dilemma situations require the existence of a state, of course. But one does not have to think very long to come up with *some* public goods with sufficiently strong joint supply and nonexclusion properties to require the participation of *all* members of a given geographic area. If such collective goods exist, then we have an explanation for a unanimous agreement to provide them.[5]

We have seen, however, how the provision of public goods is plagued by the free-rider problem; the cooperative solution to the prisoners' dilemma game is dominated. The notion of a social contract, with the connotation of mutual obligations and rewards and penalties for abiding by the contract, may serve a useful purpose in winning adherence to the provisions of the collective agreement.

Rawls is concerned throughout much of the latter part of his book with the problem of obtaining a stable, well-ordered, just society (pp. 453–504). To do so, individuals must adhere to the principles of justice incorporated in the social contract not only in the original position, but also, by and large, in daily life when they are cognizant of their actual positions. One of the important advantages claimed for the principles derived from the original position is that they stand a greater chance of compliance in the real world than any of their competitors (pp. 175–80). For this to be true, however, it is necessary that the principles be formulated so that all individuals can determine fairly readily what conduct compliance requires, and of course, all must be compelled by the nature of the arguments for compliance based on a consensus reached in the original position.

[4] For a review of this literature, see Gough (1957).

[5] For a reluctant demonstration that this is so for at least one category of public goods, see Nozick (1974).

To see that the first condition may be a problem in the Rawlsian system, consider the following example presented by Hart (1973). The application of Rawls's (pp. 201–5) first principle requires that one liberty be constrained only for the advancement of another. This requires that individuals in the original position trade off the benefits from advancing one liberty against the costs of constraining another. Private property, including the right to own land, is one of the possible freedoms that Rawls allows in his system. But the right to own land might be defined to include the right to exclude trespassers, and this in turn would conflict with the right of free movement. Thus, rights to exclude trespassers and rights to free movement are among those that would have to be sorted out at the original position. Now suppose that a farmer and a hiker get into conflict over the hiker's right to cross the farmer's field. The priority of liberty principle will do nothing to promote compliance with the social contract if the farmer and hiker, or any two people selected at random, are not likely to agree on whose right is to be preserved upon adopting the reflective frame of mind called for in the original position. But, as defined, the original position does not seem to contain enough information to allow one to sort out the priority of different liberties, and thus compliance with these important stipulations of the social contract cannot be presumed.[6]

It might be possible to resolve this kind of conflict from the original position if more information were available to individuals in this position. If they knew the amount of land available, population densities, the impact of trespassing on agricultural productivity, the alternatives to trespassing and their costs, and the like, they might be able to specify whether the right to own property took precedence or not, or even work out mixed cases in which trespassing was prohibited on land smaller than some size, but public pathways were required on larger plots. However, allowing this kind of information would in effect allow individuals to make probability calculations, and this is precluded from the original position by the characteristics of the veil of ignorance. Thus, at the level of generality at which they are derived, the principles inherent in Rawls's social contract may be an imperfect guide for compliance.

The problem of compliance can be likened to the existence of a core in a game in which individuals behind the veil of ignorance choose principles to govern the distribution of resources once the veil is lifted. If a core exists, no individual or coalition of individuals will choose to return behind the veil of ignorance and draft new principles. Howe and Roemer (1981) show that the difference principle, defined as maximizing the *incomes* of the lowest income group, yields a core to the game if all individuals are extremely risk-averse in the sense that they will join a new coalition only when they can *guarantee* themselves a higher income. Less extreme risk aversion leads to less extreme (egalitarian) principles of justice.

Rawls explicitly rejects a defense of the difference principle based on individual attitudes toward risk and similar utilitarian concepts (p. 172). Rather, he argues for greater compliance with his social contract than with a set of principles based on utilitarianism on the grounds that one could not expect compliance from the

[6] Ackerman raises similar criticisms of the problem of conceptualizing what principles the impartial or ethical observer arrives at, even assuming that one is able to assume an impartial frame of mind (1980, pp. 327–42).

poor under any set of principles requiring them to make sacrifices for the rich, as might occur under a set of utilitarian principles (pp. 175–80). But, under the difference principle, the rich are to be asked to make sacrifices (possibly quite large) for the benefit (possibly quite small) of the poor. This could lead to a problem of noncompliance by the rich.[7] Rawls (1974, p. 144) has responded to this form of criticism by noting that the "better situated . . . are, after all, more fortunate and enjoy the benefits of that fact; and insofar as they value their situation relatively in comparison with others, they give up much less." However plausible this argument is in its own right, it does not seem adequate as a part of a defense of the difference principle within the context of Rawls's theoretical framework. The latter would seem to dictate that the appeal for compliance rests on the inherent justness (fairness) of the principle's application and the proposition that the rich would agree to this principle from behind the veil of ignorance. But here we have a difficulty. The gains to the rich are excluded from consideration under alternative distributions because probability information is barred from the original position.

The exclusion of probability information cannot be defended entirely on the grounds that it would lead to principles favoring one *individual against another*. Knowing the numbers of rich and poor in the country and yet not knowing one's own income could still lead one to select a set of rules that were impartial with respect to one's own future position. But these rules would undoubtedly not include the difference principle.[8] As Rawls's three arguments in defense of the difference principle indicate, in the presence of general knowledge about probabilities something more akin to a utilitarian principle of distribution giving some weight to the interests of rich as well as poor would be selected. Rawls's chief reason for ruling out information about probabilities from the original position would thus appear to be to remove rational calculations of an average utility sort. But, as Nagel pointed out (1973, pp. 11–12), the elimination of competing principles is supposed to be a *consequence* of the working out of the justice-as-fairness concept, not a presupposition of the analysis.[9] Note also that Rawls does allow individuals in the original position certain pieces of information that are particularly favorable to the selection of his twin principles, for example, a period of moderate scarcity reigns, and individuals care little for what they receive above the base minimum. A utilitarian might ask that this information be excluded from the original position along with the general probability information that serves to handicap the selection of utilitarian rules. In any event, the construction of the arguments in favor of the difference principle is such that an individual more favorably situated than the worst-off individual in the society might question whether his interests have been fairly treated in the original position. If he does, we have a compliance problem. Rawls's social contract and his arguments in support seem to be constructed entirely for the purpose of achieving the compliance of only one group, the worst-off individuals (pp. 175–80).

[7] Nagel (1973, p. 13); Scanlon (1973, pp. 198 ff.); Klevorick (1974); Mueller, Tollison, and Willett (1974a); Nozick (1974, pp. 189–97).

[8] Nagel (1973); Mueller, Tollison, and Willett (1974a); Harsanyi (1975a).

[9] See, also, Hare (1973, pp. 90–1) and Lyons (1974, pp. 161 ff.).

Problems of compliance could also arise among the various candidates for the worst-off position (Klevorick, 1974). As Arrow (1973) and Harsanyi (1975a) have noted, these are likely to include the mentally and physically ill and handicapped as well as the very poor. But with the set of primary goods defined over several dimensions, individuals in the original position will be forced into interpersonal utility comparisons of the type Rawls seeks to avoid (Arrow, 1973; Borglin, 1982). Should individuals disagree in their rankings, then the problem of noncompliance could again arise, since those who fail to qualify as the worst off under Rawls's difference principle receive no weight whatsoever in the social outcome. If someone truly believed that the affliction he bore was the worst that anyone could possibly bear, it is difficult to see how one could make a convincing argument to him that his position was ignored in the meting out of social justice, on the grounds that from an original position, in which he did not know he had this affliction, he would weigh it below some other. He in fact has it, and the knowledge this imparts to him convinces him that he is the worst off.

Inevitably, in trying to justify an actual implementation of the difference principle and win compliance, one is led to appeal for compliance by an individual by pointing to another who is unquestionably worse off. This resembles Varian's (1974, 1976) suggestion that the difference principle should be defined in terms of envy; the worst-off individual is the one that no one envies. Here, of course, we can still have conflicts. The blind may envy those who are paralyzed but can see, and the latter may envy those who can walk but are blind. Even if the envy relationship is, from behind the veil of ignorance, transitive, the risk here is that the individual selected as the worst off will be someone who is very bad off indeed – someone perhaps like the pathetic creature in Trumbo's *When Jonny Comes Marching Home*. Literal application of the procedure to someone in this position could lead to the expenditure of immense resources to achieve a very modest improvement in individual welfare. Arrow (1973) is undoubtedly right in arguing that this is the type of special case to which Rawls's principles are not meant to apply. But the number of special cases is likely to be large, and it is particularly awkward to set aside these often pitiable and ethically difficult cases from the application of the principles of justice, because it is precisely these kinds of cases that one would like an ethical theory to handle.

These problems are all variants on the general problem of compliance raised in the example of the rich and poor. Much of Rawls's discussion of the difference principle seems to be couched in a comparison of *the* rich and *the* poor, as if there were but two groups to compare and one criterion by which to compare them. But in reality there are many possible groupings of individuals and many possible dimensions over which their welfares can be defined. Thus, a line must be drawn on the basis of some sort of interpersonal utility comparisons, around those who are to be categorized as *the* worst off. Unless a fair consensus exists on where this line is to be drawn, compliance with the principles of justice may not be forthcoming (Klevorick, 1974), for the difference principle treats all of those outside of the line, the rich and the not so rich, as being equally rich. This may lead to compliance problems among the very rich, who have to make great sacrifices for the worst off, and among the fairly poor, who receive no special treatment at all. In this way, a utilitarian principle, which weighed each individual's welfare to some degree, might

achieve greater compliance than the difference principle, which ignores the welfare of all but a single group (Harsanyi, 1975a).

25.4.2 *The two principles of justice*

Even if we accept the preceding criticisms of the social contract aspect of Rawls's theory, it is still possible to consider the two principles of justice based on the justice-as-fairness argument as candidates for a set of political institutions. The question then is, can the arguments behind these two principles be sustained?

The ethical support for these two principles is derived from the impartiality characterizing the original position and the unanimity that stems from it. Is, then, the original position truly impartial with respect to all competing principles of justice? In setting up the problem as one in which "free and equal persons" voluntarily assent to principles to govern their lives, liberty seems to receive a prominent position from the start.[10] It is perhaps no surprise, therefore, that liberty is "chosen" as the top-priority principle from the original position.

A similar argument has been made by Nozick (1974, pp. 198–9) against the difference principle: "A procedure that founds principles of distributive justice on what rational persons who know nothing about themselves or their histories would agree to *guarantees that end-state principles of justice will be taken as fundamental*" (italics in original). Given that people know nothing about the economic structure of society, about how primary goods and the other outcomes of economic and social interaction are produced, they have no choice but to ignore these intermediate steps, and any principles of justice that might govern them, and focus on final outcomes, the end distribution of primary goods. Nozick argues that this conceptualization of the setting for choosing principles of justice excludes consideration of principles that would govern the *process* of economic and social interaction. In particular, it excludes consideration of an *entitlement* principle of distributive justice, in which individuals are entitled to their holdings as long as they came to them via voluntary transfers, exchanges, and cooperative productive activity, that is, by legitimate means (Nozick, 1974, pp. 150–231). To choose such a principle, one would have to know something about how the society functions, information unavailable in the original position.

The flavor of Nagel's and Nozick's criticisms can possibly be captured by returning to our example of the rule-making card game. In this particular example, it is highly unlikely that the players choose rules to bring about particular end-state distributions. If they did, they would probably agree to have all players wind up with an equal number of chips, or points. But this would destroy much of the purpose of the game, which is presumably to match each player or couple's skill against that of the other players, given the chance distribution of the cards. The fun of the game is in the playing, and *all* of the rules would govern the *process* by which winners are selected and not the *final positions* of the winners.

My point here is not to argue that life is like a game of cards and thereby defend Nozick's entitlement theory. But it is valid to argue that individuals may want to

[10] Nagel (1973, pp. 5–11). The quoted words are from Rawls (p. 13).

consider the *context* and *process* by which outcomes are determined, perhaps along with these outcomes, in choosing principles of justice.[11] It is ironic that Rawls's theory, which derives its conception of justice from the process by which principles are chosen, rules out all consideration of principles that deal with the subsequent process of social interaction (except for those contained in the equal liberty principle) (Nozick, 1974, p. 207). Indeed, the theory based on the notion of justice as fairness seems to exclude the selection of a principle of justice that would give to each individual anything that he had acquired by fair means, a principle that does resemble Nozick's entitlement principle.

Even if we accept Rawls's constraints on the information available in the original position and view the problem as one of selecting an end-state distribution principle, it is not clear that the difference principle is the one that would necessarily be chosen. As Harsanyi (1975a) and Binmore (1994, pp. 327–33) have argued, in the absence of objective probability information, we implicitly and almost instinctively apply subjective probability estimates, or act as if we do, when making decisions. Suppose that the prize for correctly identifying the color of the ball drawn from a bag in our previous example is $5, and nothing is paid, or charged, if the color is incorrectly guessed. Since the game is free, even a person who is maximin risk-averse will play. If she chooses white, she is implicitly assuming that the probability of a white ball being chosen is equal to or greater than 0.5. If she chooses black, the reverse. If she is indifferent between the choice of color and perhaps uses a fair coin to decide, she is implicitly applying the principle of insufficient reason. It is difficult to believe that individuals in the original position will not form probability estimates of this sort, perhaps to eliminate the awkward special cases of physical and mental illness discussed above, and if they do they are unlikely to choose the maximin rule.[12]

It is also possible, under the assumptions that Rawls makes about the original position, that utilitarianism would give outcomes rather similar to those of Rawls's system.[13] To assume that "the person choosing has a conception of the good such that he cares very little, if anything, for what he might gain above the maximum stipend that he can, in fact, be sure of by following the maximin rule" is equivalent to assuming rapidly diminishing marginal utility of income (primary goods). Incorporated into von Neumann-Morgenstern utility indexes, this assumption implies extreme risk aversion and would undoubtedly lead to fairly egalitarian redistribution rules, although probably not the difference principle as long as individuals care something for what lies above the minimum. More generally, under the rather favorable economic conditions that exist when the special conception of justice, including the difference principle and the lexicographic ordering of the two principles,

[11] "The suppression of knowledge required to achieve unanimity is not equally fair to all the parties. . . . [It is] less useful in implementing views that hold a good life to be readily achievable only in certain well-defined types of social structure, or only in a society that works concertedly for the realization of certain higher human capacities and the suppression of baser ones, or only certain types of economic relations among men" (Nagel, 1973, p. 9).

[12] For additional discussion of the implausibility of the maximin criterion even under the assumptions Rawls makes, see Sen (1970a, pp. 135–41); Arrow (1973); Hare (1973); Nagel (1973); Mueller, Tollison, and Willett (1974a); Harsanyi (1975a); and Binmore (1994, pp. 315–33).

[13] Arrow (1973), Lyons (1974), and Harsanyi (1975a).

is chosen, it is likely that utilitarianism would also greatly favor liberty and substantial redistribution. Arrow (1973) points out that an additive social welfare function will order liberty lexicographically over all other wants, if all individuals do, as they might given enough wealth. Rawls's arguments that utilitarianism would produce significantly different outcomes, for example, slavery, often seem to rest on the assumption that utilitarianism is operating in the harsher economic environment under which only Rawls's *general* conception of justice applies. But this general conception of justice also allows trade-offs between liberty and economic gain and thus resembles utilitarianism to this extent (Lyons, 1974).

25.4.3 *Experimental evidence*

The critiques of the maximin principle discussed in the previous subsection revolve around the plausibility of the assumption that individuals choose this principle from behind the veil of ignorance. An alternative to merely speculating about what principle individuals *would* choose is to run experiments to see what they *do* choose.

Frohlich, Oppenheimer, and Eavey (1987) presented students with four possible redistribution rules (Rawls's rule of maximizing the floor, maximizing the average, maximizing the average subject to a floor constraint, and maximizing the average subject to a range constraint). The students were made familiar with the distributional impacts of the four rules and were given time to discuss the merits and demerits of each. In 44 experiments in which students were uncertain of their future positions in the income distribution, the five students in each experiment reached unanimous agreement on a redistributive rule to determine their final incomes. Not once did they choose Rawls's rule of maximizing the floor. The most popular rule, chosen 35 out of 44 times, was to maximize the average subject to a floor constraint. Similar experiments conducted in Canada, Poland, and the United States have all found (1) that individuals can unanimously agree on a redistributive rule, and (2) that it is almost never Rawls's maximin rule, but rather some more utilitarian rule like maximizing the mean subject to a floor (Frohlich and Oppenheimer, 1992).

Hoffman and Spitzer (1985) also found that students in an experimental setting employ a principle of distributive justice that is neither straight Rawlsian egalitarianism nor simple utilitarianism. Rather, in the context of their experiment, students employed what appeared to be a "just desserts" principle, a principle consistent with Nozick's entitlements principle. Some of the results in Frohlich and Oppenheimer's (1992, ch. 9) experiments can also be interpreted as supporting the selection of a just-desserts principle from behind a veil of ignorance.

25.5 Two utilitarian defenses of the maximin principle

25.5.1 *Maximin as a means to obtain compliance*

As we have already discussed at length, Rawls places great emphasis on the importance of including provisions in the social contract that induce subsequent compliance with it. His total rejection of utilitarian calculations is largely motivated by this goal of compliance.

Binmore (1994) recently developed a social contract theory that – like that of Rawls – places great emphasis on compliance, but follows Harsanyi in assuming that individuals are capable of making cardinal, interpersonal utility comparisons and of calculating the probabilities that they will occupy different positions once they remove the veil of ignorance. On the other hand, he also seeks to distinguish his social contract theory from their theories.

> In particular, the term *social contract* shall not be understood in the quasi-legal sense adopted, for example, by Harsanyi and Rawls. I shall emphatically *not* argue that members of society have an *a priori* obligation or duty to honor the social contract. On the contrary, it will be argued that the only viable candidates for a social contract are those agreements, explicit or implicit, that police *themselves*. Nothing enforces such a self-policing social contract beyond the enlightened self-interest of those who regard themselves as a party to it. (Binmore, 1994, p. 30, emphasis in the original)

Binmore assumes a thin veil of ignorance, which only conceals the future identities of those bargaining in the original position. Each person knows her current utility level, the utility levels of all future persons in every possible state of the world, and each person can calculate the probability that she will be any one of these future persons. Thus, all of the information needed to maximize a Harsanyi SWF is present in the original position, and rational individuals would write a social contract that achieved this end *if* the provisions of this contract could be enforced. But there is no way to enforce these provisions, and so the social contract must be written in such a way as to make it self-enforcing when people follow their enlightened self-interest (Binmore, 1994, pp. 52–3).

To illustrate the nature of Binmore's arguments, consider the following example involving a community of two. In the absence of a social contract Adam and Eve would experience utility levels of 1 and 2, respectively. By agreeing to cooperate in certain prisoners' dilemma situations three possible alternative states of the world are possible – $x(6, 8)$, $y(5, 10)$, and $z(4, 12)$ – where the first number in parentheses is the utility level experienced by Adam, the second the level experienced by Eve.[14] Because Adam and Eve are in a bargaining situation, they consider only the utility *gains* that they would experience under each possible social contract. Thus, in the absence of any form of uncertainty, they might be expected to reach the outcome predicted by Nash's (1950) solution to the bargaining problem – the outcome maximizing the Nash SWF (see ch. 23). The values for the three possible social outcomes are

$$\begin{aligned} W_N(x) &= (6-1)(8-2) = 30, \\ W_N(y) &= (5-1)(10-2) = 32, \\ W_N(z) &= (4-1)(12-2) = 30. \end{aligned} \tag{25.1}$$

[14] We are, of course, dealing with cardinal interpersonally comparable utilities here. Binmore (1994, 1998) devotes a lot of space to discussing the advantages and difficulties with these measures.

Adam and Eve would select y if they were certain of their future identities and could commit not to cheat on the terms of the contract in the future.

If Adam and Eve could commit not to cheat on the terms of the contract in the future, but assumed that they had an equal probability of being one another, they would ignore the status quo distribution and select z as it maximizes the Harsanyi SWF

$$W_H(x) = 6 + 8 = 14, \qquad W_H(y) = 5 + 10 = 15, \qquad W_H(z) = 4 + 12 = 16. \tag{25.2}$$

However, since Adam and Eve cannot commit not to cheat on the terms of the contract in the future, they select the social contract that produces x – the maximin outcome in terms of increments in utility – as the provisions of this choice, according to Binmore, are self-enforcing.

Binmore, like Rawls, assumes that the only threat to the stability of the social contract comes from the worst-off individual. Adam will not violate the contract that produces x because his gain is smaller under both alternative social contracts. But might x not be overturned by Eve, once she knows her identity, because her gain would be greater under either alternative social contract? Note that x is the maximin choice even if Eve's payoff under y is 100 or 100 million. Conceivably there might exist some utility payoff to Eve under a different state from x that would induce her to take the plunge of throwing the community back into the state of anarchy, in the hopes that a different social contract would be chosen. If this is so, then the defense of the maximin criterion as a guarantee for compliance fails. We present some other criticisms of Binmore's approach below.

25.5.2 *Maximin as a redistribution principle*

Consider now the theory of Pareto-optimal redistribution first proposed by Hochman and Rodgers (1969). Rich Mutt gives to poor Jeff because Jeff's utility is an argument in Mutt's utility function. Assuming that Jeff's utility is positively related to his income, we can write Mutt's utility as a function of both Mutt's and Jeff's incomes:

$$U_M = U(Y_M, Y_J). \tag{25.3}$$

Given such a utility function, we can expect rich Mutt to make voluntary transfers to poor Jeff if the latter figures heavily enough in Mutt's utility function. In a world of more than one Jeff, Mutt will receive the highest marginal utility from giving a dollar to the poorest Jeff. Thus, although the Pareto-optimal approach to redistribution does not fully justify the maximin principle, it does justify a redistribution policy that focuses sole attention on the worst-off individual or group (von Furstenberg and Mueller, 1971). An altruistic utilitarian and a Rawlsian will both consider the welfare of only the worst-off individual(s) in society.[15]

[15] For two additional utilitarian defenses of the difference principle, see Buchanan (1976) and Chu and Liu (1998).

25.6 The social contract as a constitution

From a public choice perspective what is of most interest from the literature on social contracts is the potential insights it may yield for the design of political institutions. If we agree with Rawls that the basic institutions of society, including its political institutions, *ought* to be selected from behind a veil of ignorance, what should those institutions look like? What do Rawls's two principles of justice imply regarding the optimal design of a constitution?

The implications of the first principle seem fairly clear. The social contract, and by logical extension the constitution, should protect liberty. Rights to free speech, privacy, and the like come readily to mind as political embodiments of Rawls's equal liberty principle.

What about the second principle? What set of political institutions would embody the intent of the difference principle? This principle has quite obvious implications for the distribution of income and wealth or, as Rawls prefers, of primary goods, and much of the discussion of it by both Rawls and Binmore can be most easily understood in this context. Distributional questions are not the only issues a society must resolve, however. What are the implications of the difference principle or the maximin criterion for the provision of public goods, or for the resolution of conflict issues that do not involve the distribution of income or primary goods? What electoral and voting rules does the difference principle imply?

The most straightforward application of the difference principle to choosing a voting rule to decide public goods issues suggests that the unanimity rule should be chosen. Who is the worst-off person when a public good is provided using a qualified majority rule – one of the persons who votes against its provision. Because it is always possible to determine a set of tax shares and public good quantity so that every person is made better off, maximizing the welfare of the worst-off individual would seem to require a continual reformulation of the issue until a set of tax shares and a public good quantity is found to which all agree. But against this interpretation of the difference principle all of the objections that have been made against the unanimity rule can be levied.

Consider next a simple, conflictual issue – the maximum speed to be allowed on the public highways. Who is harmed by high speed limits? – those injured in accidents with the worst-off person clearly being someone killed in a road accident. What speed limit would maximize the welfare of the worst off individual? – a limit so low that it precluded any serious accidents. The extreme aversion to risk that characterizes the maximin criterion as a principle for making choices in the face of uncertainty is readily apparent here, as is the impracticality of applying it either to specific issues of conflict or as a guide for choosing a voting rule to resolve such issues.

The same can be said of Binmore's application of the maximin criterion. Recall that Binmore's derivation of the optimality of this criterion does not hinge on individual attitudes toward risk, but rather stems from his concern for avoiding defections from the social contract's provisions, once the veil of ignorance is lifted. This goal would also seem to imply the application of the unanimity rule in the postagreement stage. Who are the most likely people to defect from a decision made under some

alternative, qualified majority rule? – the losers under this rule, those "tyrannized" by the application of a less-than-unanimity rule, when unanimity is possible. Who are the most likely to defect from a decision to allow cars to travel at speeds that produce some fatal accidents? – those who do not own cars and experience only the costs from such a decision. To avoid the possibility of future defections from collective decisions, one must avoid creating losers from these decisions, and this implies using the unanimity rule whenever consensus is possible.

These observations are not made as criticisms of the social contract theories of Rawls and Binmore, for these theories were not intended to produce principles that allow society to set speed limits or even to choose a voting rule to determine speed limits. But these examples make clear that these social contract theories are not going to be of much help in making these more mundane collective choices. Indeed, when Rawls comes to discuss why the simple majority rule would be included in a constitution written in accordance with his principles of justice, he *does not* demonstrate how this rule follows logically from these principles. Instead, Rawls assumes that all citizens and legislators have already joined the just social contract, so that "legislative discussion must be conceived not as a contest between interests, but as an attempt to find the best policy as defined by the principles of justice. I suppose, then, as part of the theory of justice, that an impartial legislator's only desire is to make the correct decision in this regard" (p. 357). If all legislators were fully informed, they would all know what the correct decision is, and the unanimity rule could be used. The only reason not to use it is that legislators are not fully informed. Therefore Rawls opts for the simple majority rule using Condorcet's original defense of this rule. It is used as a sampling procedure to aggregate the views of the impartial legislators and thereby to obtain "a best judgment" as to what the correct decision is (pp. 357–8).[16]

I assume that very few readers who have come this far with me will share Rawls's belief that legislators are impartial seekers of the correct decisions for the community, and that the only task of politics is to sort out what these correct decisions are.[17] We need to consider the political institutions to be chosen from behind the veil of ignorance under the assumption that politics *is* a "contest between interests," and entertain the possibility that these interests are narrowly defined. This was, in fact, the exercise that Buchanan and Tullock (1962) set for themselves when writing *The Calculus of Consent*. We take it up in Chapter 26.

Bibliographical notes

Daniels (1974) contains an excellent set of papers analyzing and criticizing Rawls. (Page references in this chapter are to the reprinted versions in Daniels.) The books by Nozick (1974), Wriglesworth (1985), Gauthier (1986), and Barry (1989) can be

[16] Condorcet's arguments are reviewed in ch. 6.

[17] Not surprisingly, Rawls does not regard public choice as an appropriate methodology for determining the optimal design of just political institutions, as the following statement reveals: "the application of economic theory to the actual constitutional process has grave limitations insofar as political theory is affected by men's sense of justice" (p. 360).

said to have been inspired by Rawls's theory. Binmore's (1994, 1998) two-volume treatise links modern game theory to the classical social contract theory from Hobbes on up to Rawls. It also shows the relationship between this work and the SWF of Harsanyi. It contains an exhaustive discussion of von Neumann-Morgenstern utility indexes and of cardinal, interpersonal utility comparisons. John Rawls's recent thoughts on social justice are presented in his 1999 book.