The original position is a central feature of John Rawls's social contract account of justice, “justice as fairness,” set forth in *A Theory of Justice* (TJ). It is designed to be a fair and impartial point of view that is to be adopted in our reasoning about fundamental principles of justice. In taking up this point of view, we are to imagine ourselves in the position of free and equal persons who jointly agree upon and commit themselves to principles of social and political justice. The main distinguishing feature of the original position is “the veil of ignorance”: to insure impartiality of judgment, the parties are deprived of all knowledge of their personal characteristics and social and historical circumstances. They do know of certain fundamental interests they all have, plus general facts about psychology, economics, biology, and other social and natural sciences. The parties in the original position are presented with a list of the main conceptions of justice drawn from the tradition of social and political philosophy, and are assigned the task of choosing from among these alternatives the conception of justice that best advances their interests in establishing conditions that enable them to effectively pursue their final ends and fundamental interests. Rawls contends that the most rational choice for the parties in the original position are two principles of justice: The first guarantees the equal basic rights and liberties needed to secure the fundamental interests of free and equal citizens and to pursue a wide range of conceptions of the good. The second principle provides fair equality of educational and employment opportunities enabling all to fairly compete for powers and positions of office; and it secures for all a guaranteed minimum of all-purpose means (including income and wealth) individuals need to pursue their interests and to maintain their self-respect as free and equal persons.

- 1. Historical Background: the Moral Point of View
- 2. The Original Position and Social Contract Doctrine
- 3. The Veil of Ignorance
- 4. Description of the Parties: Rationality and the Primary Social Goods
- 5. Other Conditions on Choice in the Original Position
  - 5.1 The Circumstances of Justice (TJ §22)
  - 5.2 Publicity and other Formal Constraints of Right (TJ §23)
  - 5.3 The Stability Requirement
- 6. The Arguments for the Principles of Justice from the Original Position
  - 6.1 The Argument from the Maximin Criterion (TJ sects. 26–28)
  - 6.2 The Strains of Commitment
  - 6.3 Stability, Publicity, and Self-Respect
  - 6.4 The Argument for the Difference Principle
- 7. Is the Original Position Necessary or Relevant?
- 8. The Original Position and the Law of Peoples
- 9. Constructivism, Objectivity, Autonomy, and the Original Position
- Bibliography
  - Primary Sources
  - Secondary Sources
- Academic Tools
- Other Internet Resources
- Related Entries
1. Historical Background: the Moral Point of View

The idea of a moral point of view can be traced back to David Hume's account of the “judicious spectator.” Hume sought to explain how moral judgments of approval and disapproval are possible given that people normally are focused on achieving their particular interests. He conjectured that in making moral judgments individuals abstract in imagination from their own particular interests and adopt an impartial point of view from which they assess the effects of others' actions on the interests of everyone. Since, according to Hume, we all can adopt this impartial perspective in imagination, it accounts for our agreement in moral judgments. (See Hume, 1739/1978, 581; Rawls, LHMP 84–93, LHPP 184–187.)

Subsequently, philosophers posited similar perspectives for moral reasoning designed to yield impartial judgments once individuals abstract from their particular aims and interests and assess situations from an impartial point of view. But rather than being mainly explanatory of moral judgments like Hume's “judicious spectator,” the role of these impartial perspectives is to serve as a basis from which to assess and justify moral rules or principles. Kant's categorical imperative procedure, Adam Smith's “impartial spectator,” Rousseau's general will, and Sidgwick's “point of view of the universe” are all different versions of such a moral point of view.

An important feature of the moral point of view is that it is designed to represent something essential to the activity of moral reasoning. For example, Kant's categorical imperative is envisioned as a point of view any reasonable person can adopt in deliberating about what he/she ought morally to do (Rawls, CP 498ff). When joined with the common assumption that the totality of moral reasons is final and override non-moral reasons, the moral point of view might be regarded as the most fundamental perspective that we can adopt in practical reasoning about what we ought to do.

Rawls's idea of the original position, as initially conceived, is his account of the moral point of view with regard to matters of justice. The original position is a hypothetical perspective that we can adopt in our moral reasoning about the most basic principles of social and political justice. What primarily distinguishes Rawls's impartial perspective from its antecedents (in Hume, Smith, Kant, etc.) is that, rather than representing the judgment of one person, it is conceived socially, as a general agreement by (representatives of all adult) members of an ongoing society. The point of view of justice is then represented as a general social “contract” or agreement, similar to Rousseau's social contract.

2. The Original Position and Social Contract Doctrine

Historically the idea of a social contract had a more limited role than Rawls assigns to it. In Hobbes and Locke the social contract serves as an argument for the legitimacy of political authority. Hobbes argues that in a pre-social state of nature it would be rational for all to agree to authorize one person to exercise the absolute political power needed to enforce norms necessary for social cooperation. By contrast, Locke argued against absolute monarchy by contending that no existing political constitution is legitimate or just unless it could be contracted into starting from a position of equal right within a (relatively peaceful) state of nature, and without violating any natural rights or duties. For Rousseau and perhaps Kant too, the idea of a social contract played a different role: as part of their accounts of the General Will, the social contract is a point of view that lawmakers and citizens should adopt for assessing existing laws deciding on measures that achieve justice and citizens' common good. Rawls generalizes on Locke's, Rousseau's and Kant's natural right theories of the social contract (TJ vii/xviii): the purpose of his original position is to yield principles to determine and assess the justice of political constitutions and of economic and social arrangements. To do so, he seeks in the original position “to combine into one conception the totality of conditions which we are ready upon due reflection to recognize as reasonable in our conduct towards one another” (TJ 587/514).
Why does Rawls represent principles of justice as originating in a kind of social contract? (Later we’ll consider the objection that the original position is not really social, but is the rational choice of one representative person.) Rawls says that “justice as fairness assigns a certain primacy to the social” (CP 339). Unlike Kant's categorical imperative procedure, the original position is designed to represent the predominantly social bases of justice. To say that justice is predominantly social does not mean that people do not have “natural” moral rights and duties outside society or in non-cooperative circumstances—Rawls clearly thinks there are certain human rights (see LP sect.10) and natural duties (TJ sects. 19, 51) that apply to all human beings as such. But whatever our natural or human rights and duties may be, they do not provide an adequate basis for ascertaining the rights and duties of justice that we owe one another as members of the same ongoing political society. It is in large part due to “the profoundly social nature of human relationships” (PL 259) that Rawls sees political and economic justice as grounded in social cooperation and its reciprocity. For this reason Rawls eschews the idea of a state of nature wherein pre-social but fully rational individuals agree to cooperative norms (as in Hobbesian views), or where pre-political persons with antecedent natural rights agree on the form of a political constitution (as in Locke). We are social beings in the sense that in the absence of society and social development we have but inchoate and unrealized capacities, including our capacities for rationality, morality, even language itself. As Rousseau says, outside society we are but “stupid and shortsighted animals” (Rousseau, bk.I, ch.8, par. 1). This undermines the main point of the idea of a state of nature in many views, which is to distinguish the rights, claims, duties, powers and competencies we have prior to membership in society from those we acquire as members of society. Not being members of some society is not an option for us. In so far as we are rational and reasonable beings at all, we have developed as members of a society, within its social framework and institutions. The traditional idea of pre-social or even pre-political rational moral agents thus plays no role in Rawls's account of justice and the social contract; for him the state of nature is an idea without moral significance (PL 278–280). The original position is set forth largely as an alternative to the state of nature and is regarded by Rawls as the appropriate initial situation for a social contract. (Below we consider a further reason behind Rawls's rejection of the state of nature: it does not adequately allow for impartial judgment and the equality of persons.)

Another way in which Rawls represents the “profoundly social” bases of principles of justice is by focusing on “the basic structure of society.” The “first subject of justice,” Rawls says, is principles that regulate the basic social institutions that constitute the “basic structure of society” (TJ sect.2). These basic institutions include the political constitution, which specifies procedures for legislating and enforcing laws and the system of trials for adjudicating disputes; the bases of the economic system, including the norms of property, its transfer and distribution, contractual relations, etc. which are all necessary for economic production, exchange, and consumption; and finally norms that define and regulate permissible forms of the family, which is needed to perpetuate society. It is the role of principles of justice to specify and assess the system of rules that constitute these basic institutions, and determine the fair distribution of rights, duties, opportunities, powers and positions of office to be realized within them. What makes these institutions and their arrangement the first subject for principles of social justice is that they are all necessary to social cooperation and have such profound influences on our circumstances, aims, characters, and future prospects. No society could exist without certain rules of property, contract, and transfer of goods and resources, for they make economic production, trade, and consumption possible. Nor could a society long endure without some political mechanism for resolving disputes and making, revising, interpreting, and enforcing its economic and other cooperative norms; or without some form of the family, to reproduce, sustain, and nurture members of its future generations. This is what distinguishes the social institutions constituting the basic structure from other profoundly influential social institutions, such as religion; religion and other social institutions are not basic in Rawls's sense because they are not generally necessary to social cooperation among members of society. (Even if they are ideologically necessary to sustain particular societies, many societies can and do exist without the support of religious institutions).

Another reason Rawls regards the original position as the appropriate setting for a social contract is implicit in his stated aim in A Theory of Justice: it is to discover the most appropriate moral conception of justice for a
democratic society wherein persons regard themselves as free and equal citizens (TJ viii/xviii). Here he assumes an ideal of citizens as “moral persons” who regard themselves as free and equal, have a conception of their rational good, and also have a “sense of justice.” “Moral persons” (an 18th century term) are not necessarily morally good persons, but instead are capable of being rational in that they have the capacities to form, revise and pursue a conception of the good; and also they are capable of being reasonable since they have a moral capacity for a sense of justice—to cooperate with others on terms that are fair and to understand, apply, and act upon principles of justice and their requirements. Because people have these capacities, or “moral powers,” we hold them responsible for their actions, and they are regarded as capable of pursuing their interests and engaging in social cooperation. Rawls’s idea is that, being reasonable and rational, persons (like us) who regard ourselves as free and equal should be in a position to accept and endorse as morally justifiable the principles of justice regulating our basic social institutions and individual conduct. Otherwise our conduct is coerced for reasons we cannot (reasonably or rationally) accept and we are not fundamentally free persons. Starting from these assumptions, Rawls construes the moral point of view from which to decide moral principles of justice as a social contract in which (representatives of) free and equal persons are given the task of coming to an agreement on principles of justice that are to regulate their social and political relations in perpetuity. How otherwise, Rawls contends, should we represent the justification of principles of justice for free and equal persons who have different conceptions of their good, as well as different religious, philosophical, and moral views? There is no commonly accepted moral or religious authority or doctrine to which they could appeal in order to discover principles of justice that all could agree to and accept. Rawls contends that, since his aim is to discover a conception of justice appropriate for a democratic society that is justifiable to its free and equal citizens and which all can endorse and accept, the appropriate way to justify this conception is by (imagining) an agreement among free and equal moral persons themselves.

How is this social contract to be conceived? It is not an historical event that must actually take place at some point in time (TJ 120/104). It is rather a hypothetical situation, a kind of thought experiment, that is designed to uncover the most reasonable principles of justice. Rawls maintains (in LHPP, cf. p.15) that the major advocates of social contract doctrine—Hobbes, Locke, Rousseau, and Kant—all regarded the social contract, as a hypothetical event. Hobbes and Locke thus posited a hypothetical state of nature in which there is no political authority, and where people are regarded as rational and (for Locke) also reasonable. The purpose of this hypothetical social contract is to demonstrate what types of governments are politically legitimate, and determine the nature of individuals' political obligations (ibid. p.16). The presumption is that if a government could or would be agreed to by all rational persons subject to it in an appropriately described pre-political situation, then it is acceptable to rational persons generally, including you and me, and hence is legitimate and is the source of our political obligations. Thus Hobbes argues that all rational persons in a state of nature would agree to authorize an absolute sovereign, while Locke comes to the opposite conclusion, contending that absolutism would be rejected in favor of constitutional monarchy. Similarly, in Rousseau and Kant, the social contract is a way to reason about the General Will, or the laws that hypothetical moral agents would all agree to in order to promote the common good and realize the freedom and equality of citizens.

Rawls employs the idea of a hypothetical social contract for more general purposes than his predecessors. He aims to provide principles of justice that can be applied to determine not only the justice of political constitutions and the laws, but also the justice of social and economic arrangements in the distribution of income and wealth, as well as educational and work opportunities, powers and positions of office and responsibility.

Some have objected that hypothetical agreements cannot bind or obligate people; only actual contracts or agreements can impose obligations and commitments (Dworkin, 1977, 150ff). In response, Rawls says that the OP is not intended to impose new obligations on us, but is to be used “to help us work out what we now think” (CP 402); it incorporates “conditions…we do in fact accept” (TJ 587/514) and is a kind of “thought experiment for the purpose of public- and self-clarification” (JF, p.17). Hypothetical agreement in the original
position does not then bind anyone to duties or commitments he/she does not already have. Its point rather is to explicate the requirements of our moral concepts of justice and enable us to draw the consequences of considered moral convictions of justice that we all presumably share. Whether we in turn consciously accept or agree to these consequences and the principles and duties they implicate once brought to our awareness is irrelevant to their justification. The point rather of conjecturing the outcome of a hypothetical agreement is that, assuming that the premises underlying the original position correctly represent our most deeply held considered moral convictions and concepts of justice, then we are committed to endorsing the resulting principles and duties whether or not we actually accept or agree to them. Not to do so implies a failure to live up to the consequences of our own moral convictions about justice.

3. The Veil of Ignorance

Rawls calls his conception “justice as fairness.” His aim in designing the original position is to describe an agreement situation that is fair among all the parties to the hypothetical social contract. He assumes that if the parties to the social contract are fairly situated and take all relevant information into account, then the principles they would agree to are also fair. The fairness of the original agreement situation transfers to the principles everyone agrees to, and further that whatever laws or institutions are required by the principles of justice are also fair. The principles of justice chosen in the original position are in this way the result of a choice procedure designed to “incorporate pure procedural justice at the highest level” (CP, 310, cf. TJ 120/104).

There are different ways to define a fair agreement situation depending on the purpose of the agreement and the description of the parties to it. For example, certain facts are relevant to entering into a fair employment contract—a prospective employee's talents, skills, experience and motivation for example— that may not be relevant to other fair agreements. What is a fair agreement situation among free and equal persons when the purpose of the agreement is fundamental principles of justice for the basic structure of society? Here it is helpful to compare Rawls's and Locke's social contracts. A feature of Locke's social contract is that it transpires in a state of nature among free and equal persons who know everything about themselves that you and I know about ourselves and each other. Thus Locke's parties know their natural talents and other personal characteristics; their racial and ethnic group, social class and occupations; their level of wealth and income, their religious and moral beliefs, an so on. Given this knowledge, Locke assumes that, while starting from a position of equal political right, the great majority of free and equal persons in a state of nature (all women and all men who do not meet a rigid property qualification) could and most likely would rationally agree to alienate their natural rights of equal political jurisdiction in order to gain the benefits of political society. Thus Locke envisions as legitimate a constitutional monarchy that is in effect a class state, a state wherein a small class of amply propertied males exercise political rights to vote, hold office, exercise political and social influence, and enjoy other important benefits and responsibilities to the exclusion of everyone else. (See Rawls, LHPP, 138–139.)

The problem with this arrangement, of course, is that gender and lack of wealth are, like absence of religious belief, not good reasons for depriving people of their equal political rights or opportunities to occupy social and political positions. These reasons are not morally relevant for deciding who qualifies to vote, hold office, and actively participate in governing and administering society. Rawls suggests that the reason Locke's social contract results in this unacceptable outcome is that it transpires (hypothetically) under unfair conditions of a state of nature, where the parties have complete knowledge of their characteristics and situations—their gender, wealth, social class, talents and skills, religious convictions, etc. Socially powerful and wealthy parties then can rely on knowledge of their “threat advantage” to extract favorable terms from those in less favorable positions (JF 16). Consequently the parties' judgments are biased by their knowledge of their circumstances and are insufficiently impartial.
The remedy for such biases of judgment is to redefine the initial situation. Rather than a state of nature Rawls situates the parties to his social contract so that they do not have access to factual knowledge that can distort their judgments and result in unfair principles. Rawls's original position is an initial situation wherein the parties are without information that enables them to tailor principles of justice favorable to their personal circumstances. “Among the essential features of this situation is that no one knows his place in society, his class position or social status, nor does any one know his fortune in the distribution of natural assets and abilities, his intelligence, strength and the like. We shall even assume that the parties do not know their conceptions of the good or their special psychological propensities. The principles of justice are chosen behind a veil of ignorance” (TJ 12/11). This veil of ignorance deprives the parties of all knowledge of particular facts about themselves, about one another, and even about their society and its history.

The parties are not however completely ignorant of facts. They know all kinds of general facts about persons and societies, including knowledge of the relatively uncontroversial laws and generalizations derivable from economics, psychology, political science, and biology and other natural sciences. They know then about the general tendencies of human behavior and psychological development, about biological evolution, and about how economic markets work, including neo-classical price theory of supply and demand. As discussed below, they also know about the circumstances of justice—moderate scarcity and limited altruism—as well as the desirability of the “primary social goods” that are needed by anyone to live a good life and to develop their “moral powers” and other capacities. What the parties lack however is knowledge of any particular facts about their own and other persons' lives, as well as knowledge of any historical facts about their society and its population, its level of wealth and resources, religious institutions, etc.. Rawls thinks that since the parties are required to come to an agreement on objective principles that supply universal standards of justice applying across all societies, knowledge of particular and historical facts about any person or society is morally irrelevant and potentially prejudicial to their decision.

Another reason Rawls gives for such a “thick” veil of ignorance is that it is designed to be a strict “position of equality” (TJ 12/11) that represents persons purely in their capacity as free and equal moral persons. The parties in the original position do not know any particular facts about themselves or society; they all have the same general information made available to them. They are then situated equally in a very strong way, “symmetrically” (JF 18) and purely as free and equal moral persons. They know only characteristics and interests they have in their capacity as moral persons—their “higher-order interests” in developing the moral powers of justice and rationality, their need for the primary social goods, and so on. The moral powers, Rawls contends, are the “basis of equality, the features of human beings in virtue of which they are to be treated in accordance with the principles of justice” (TJ, 504/441). Knowledge of the moral powers and their essential role in social cooperation, along with knowledge of other general facts, is all that is morally relevant, Rawls believes, to a decision on principles of justice that are to reflect people's status as free and equal moral persons. A thick veil of ignorance thus is designed to represent the equality of persons purely as moral persons, and not in any other contingent capacity or social role. In this regard the veil interprets the Kantian idea of equality as equal respect for moral persons (cf. CP 255).

Many criticisms have been leveled against Rawls's veil of ignorance. Among the most frequent is that choice in the original position is indeterminate (Sen, 2009, 11–12, 56–58). Among other reasons for this, it is said that the parties are deprived of so much information about themselves that they are psychologically incapable of making a choice, or they are incapable of making a rational choice. For how can we make any rational choice without knowing our primary ends, or fundamental values and commitments? (MacIntyre, 1981; Sandel 1982)

The answer is that we do not need to know everything about ourselves to make rational decisions about the background conditions needed to pursue our primary purposes. For example, whatever our ends, we know that personal security and an absence of social chaos are conditions of most anyone's living a good life (as Hobbes contends). Similarly, though Rawls's parties do not know their particular values and commitments,
they do know that they require an adequate share of primary social goods (rights and liberties, powers and opportunities, and income and wealth) to effectively pursue their purposes, whatever they may be. They also know they have a higher-order interest in adequately developing their moral powers, the conditions of responsible agency and social cooperation. Rawls contends that knowledge of these basic social needs (he also calls them “essential goods”) is sufficient for a rational choice by the parties in the original position. For critics to argue that they are not sufficient requires that they point to specific problems with Rawls's detailed arguments for his principles from the original position.

To the objection that choice behind the veil of ignorance is psychologically impossible, Rawls says that it is important not to get too caught up in the theoretical fiction of the original position, as if it were some historical event among real people who are being asked to do something impossible. The original position is not supposed to be realistic, but is a “device of representation” (PL 27), or a “thought experiment,” (JF, 83), that is designed to organize our considered convictions of justice and clarify their implications. The parties in it are not real but are “artificial persons” who have a role to play in this thought experiment. They represent an ideal of free and equal rational moral persons that Rawls assumes is implicit in our reasoning about justice. The veil of ignorance is a representation of the kinds of reasons and information that are relevant to a decision on principles of justice for the basic structure of a society of free and equal moral persons (TJ 17/16). Many different kinds of reasons and facts are not morally relevant to that kind of decision (e.g., information about people's race, gender, religious affiliation, wealth, and even, Rawls says more controversially, their conceptions of their good), just as many different kinds of reasons and facts are irrelevant to mathematicians' ability to work out the formal proof of a theorem. As a mathematician, scientist, or musician exercise their expertise by ignoring their knowledge of particular facts about themselves, we can do so too in reasoning about principles of justice. Rawls says we can “enter the original position at any time simply by reasoning in accordance with the enumerated restrictions on information,” PL 27) and taking into account general facts about persons, their needs, and social and economic cooperation that are provided to the parties. (TJ 120/104, 587/514)

A related criticism of Rawls's “thick” veil of ignorance is that even if the parties can make certain rational decisions in their interest without knowledge of their final ends, still they cannot come to a decision about principles of justice. For justice consists, allegedly, of the measures that effectively promote good consequences. Without knowledge what is ultimately good (however that is to be defined) the parties cannot discover the principles of justice that best promote it. This criticism is mirrored in utilitarian versions of the moral point of view, which incorporate a “thin” veil of ignorance that represents a different idea of impartiality. The impartial sympathetic observer found in David Hume and Adam Smith, or the self-interested rational chooser in John Harsanyi's utilitarian account, all have complete knowledge of everyone's desires, interests and purposes as well as knowledge of particular facts about people and their historical situations. Impartiality is achieved by depriving the impartial observer or rational chooser of any knowledge of its own identity. This leads it to give equal consideration to everyone's desires and interests, and impartially takes everyone's desires and interests into account. Since rationality is presumed to involve maximizing something, the impartial observer/chooser rationally chooses the rule or actions that maximizes the satisfaction of desires, or utility (aggregate or average), summed across all persons.

Rawls's original position with its “thick” veil of ignorance represents a different conception of impartiality than the utilitarian requirement that equal consideration be given to everyone's desires. The original position abstracts from all information about current circumstances and the status quo, including everyone's desires and particular interests. Utilitarians take peoples' desires and interests as given and seek to maximize their satisfaction; in so doing utilitarians suspend judgment regarding the moral permissibility of peoples' desires and preferences and of the social circumstances and institutions within which desires and preferences are formed. For Rawls, a primary reason for a thick veil of ignorance is to enable an unbiased assessment of the justice of existing social and political institutions and of existing desires, preferences and conceptions of the good. If the parties to Rawls's original position had knowledge of peoples' beliefs and desires, as well as
knowledge of the laws, institutions and circumstances of their society, then this knowledge would influence their decisions on principles of justice. The principles agreed to would then not be sufficiently detached from the very desires, circumstances, and institutions these principles are to be used to critically assess. Since utilitarians take peoples' desires and interests as given under existing circumstances, any principles, laws or institutions chosen behind their thin veil of ignorance will reflect and be biased by the status quo. To take an obvious counterexample, there is little if any justice in laws approved from a utilitarian impartial perspective when these laws take into account racially prejudiced preferences which are cultivated by grossly unequal, racially discriminatory and segregated social conditions. To impartially give equal consideration to everyone's desires formed under such under unjust conditions is hardly sufficient to meet requirements of justice. This illustrates some of the reasons for a “thick” as opposed to a “thin” veil of ignorance.

4. Description of the Parties: Rationality and the Primary Social Goods

Rawls says that in the original position, “the Reasonable frames the Rational” (CP 319). He means the OP is a situation where rational choice of the parties is made subject to reasonable (i.e. moral) constraints. In what sense are the parties and their choice rational? Philosophers have different understandings of practical rationality. Rawls seeks to incorporate a relatively uncontroversial account of rationality into the original position, one that he thinks most any account of practical rationality would endorse as at least necessary for rational decision. The parties are then described as rational in a formal or “thin” sense that is characteristic of the theories of rational and social choice. They are resourceful, take effective means to their ends, and seek to make their preferences consistent. They also take the course of action that is more likely to achieve their ends (other things being equal). And they choose courses of action that satisfy more rather than fewer of their purposes. Rawls calls these principles of rational choice the “counting principles” (TJ sect. 25, JF 87).

More generally, for Rawls rational persons upon reflection can formulate a conception of their good, or of their primary values and purposes and the best way of life for themselves to live given their purposes. This conception incorporates their primary aims, ambitions, and commitments to others, and is informed by the conscientious moral, religious, and philosophical convictions that give meaning for them to their lives. Ideally, rational persons have carefully thought about these things and their relative importance, and they can coherently order their purposes and commitments into a “rational plan of life,” which extends over their lifetimes (TJ sect. 63). For Rawls, rational persons regard life as a whole, and do not give preference to any particular period of it. Rather in drawing up their rational plans, they are equally concerned with their good at each part of their lives. In this regard, rational persons are prudent—they care for their future good, and while they may discount the importance of future purposes based on probability assessments, they do not discount the achievement of their future purposes simply because they are in the future (TJ, sect. 45).

These primary aims, convictions, ambitions, and commitments are among the primary motivations of the parties in the original position. The parties want to provide favorable conditions for the pursuit of the various elements of the rational plan of life that defines a good life for them. This is ultimately what the parties are trying to accomplish in their choice of principles of justice. In this sense they are rational.

Rawls says the parties in the original position are “mutually disinterested,” in the sense that they take no interest in one another's interests. This does not mean that they are generally self-interested or selfish persons, indifferent to the welfare of others. Most people are concerned, not just with their own happiness or welfare, but with that of others as well, and have all kinds of commitments to others, including other-regarding and beneficent purposes, that are part of their conception of the good. But in the original position itself the parties are not altruistically motivated to benefit each other, in their capacity as contracting parties. They try to do as best as they can for themselves and for those persons and causes that they care for. Their situation is
comparable, Rawls says, to that of trustees acting to promote the interests of the beneficiaries they represent. Trustees cannot sacrifice the well-being of their beneficiaries to benefit other trustees. If they did, they would be derelict in their duties. It is perhaps to address the common criticism that the parties to the original position are self-interested that Rawls in his later writings increasingly says that we should imagine that the parties are representatives or trustees of other individuals' interests, and seek to do as best as they can for the particular individuals that each trustee represents. In either case, Rawls believes this account of the parties' motivations promotes greater clarity, and that to attribute to the parties moral motivations or benevolence towards each other would not result in definite choice of a conception of justice (TJ, 148–9/128–9; 584/512). (For example, how much benevolence should they have towards one another? Surely not impartial benevolence towards everyone, for then we might as well dispense with the social contract and rely on an impartial spectator point of view. But if not equal concern for all parties, then how much?)

Mutual disinterest of the parties also means they are not moved by envy or rancor towards each other or others generally. This implies that the parties do not strive to be wealthier or better off than others for its own sake, and thus do not sacrifice advantages to prevent others from having more than they do. Instead, each party in the original position is motivated to do as well as he/she can in absolute terms in promoting the optimal achievement of the many purposes that make up a conception of the good, without regard to how much or how little others may have. For this reason they strive to guarantee themselves a share of primary social goods sufficient to enable them to pursue their (unknown) conception of the good.

Though the parties are not motivated by beneficence or even a concern for justice, still they have a capacity for reasonableness and a sense of justice (TJ, 145/125 rev.). Rawls distinguishes between the requirements of rationality and reasonableness; both are part of practical reasoning about what we ought to do (JF 6–7; 81–2). The concept of “the Rational” concerns a person's good—hence Rawls refers to his account of the good as “goodness as rationality.” A person's good for Rawls is the rational plan of life he/she would choose under hypothetical conditions of “deliberative rationality,” where there is full knowledge of one's circumstances, capacities, and interests, as well as knowledge of the likelihoods of succeeding at alternative life plans one may be attracted to (TJ, sect. 64). “The Reasonable” on the other hand has to do with the concept of right, which includes individual moral duties and moral requirements of right and justice applying to institutions and society. Both rationality and reasonableness are independent aspects of practical reason for Rawls. They are independent in that Rawls, unlike Hobbes, does not regard justice and the reasonable as simply the principles of prudence that are beneficial for a person to comply with in order to successfully pursue his or her purposes in social contexts. Unlike Hobbes, Rawls does not claim that an immoral person is irrational, or that morality is necessarily required by rationality. However, a rational person who violates reasonable demands of justice is unreasonable in that he or she infringes upon moral requirements of practical reasoning. Being reasonable, even if not required by rationality, is still an independent aspect of practical reason. Rawls resembles Kant in this regard (PL 25n); his distinction between the reasonable and rational parallels Kant's distinction between categorical and hypothetical imperatives.

Essential to being reasonable is having a sense of justice. The sense of justice is a normally effective desire to comply with duties and obligations required by justice; it includes a willingness to cooperate with others on terms that are fair and that reasonable persons can accept and endorse. Rawls sees a sense of justice as an attribute people normally have; it “would appear to be a condition for human sociability” (TJ, 495/433 rev.). He rejects the idea that people are motivated only by self-interest in all that they do; he also rejects the Hobbesian assumption that a willingness to do justice must be grounded in self-interest. It is essential to Rawls's argument for the feasibility and stability of justice as fairness that the parties upon entering society have an effective sense of justice, and that, as members of society, they are capable of doing what justice requires of them, either for its own sake, or because the believe this is what morality requires of them. An amoralist, Rawls believes, is largely a philosophical construct; the amoralists who actually exist Rawls sees as sociopaths.
Subsequent to *A Theory of Justice*, in Kantian and Political Constructivism, Rawls says that the parties to the original position have a “higher-order interest” in the exercise and development of their capacity for a sense of justice (as well as in their capacity to be rational), and that fulfilling this interest is one of the main aims behind their agreement on principles of justice. The parties' interest in developing these two “moral powers” is a substantive feature of Rawls's account of the rationality of free and equal persons. (In this regard, his account of goodness as rationality is not as “thin” as in social theory; cf. TJ 143/124.) Here is it is important to see that Rawls is still not attributing specifically moral motives—a desire to be reasonable and do what is right and just for their own sake—to the parties in the original position. The idea behind the parties' rationality in cultivating their sense of justice is that, since reasonableness is a condition of human sociability and social cooperation, then it is in people's rational interest—part of their good—that they normally develop their capacities for justice under social conditions. Otherwise they will not be in a position to cooperate with others and benefit from social life. A person who is without a sense of justice is wholly unreasonable and as a result is normally eschewed by others, for he or she is not trustworthy or reliable or even safe to interact with. Since having a sense of justice is a condition of taking part in social cooperation, the parties have a “higher-order interest” in establishing conditions for the full development and effective exercise of their sense of justice. The parties' interest in developing their sense of justice is then a rational interest in being reasonable; justice is then regarded *by the parties* as instrumental to their realizing their conception of the good. (Here again, it is important to distinguish the purely rational motivation of the parties or their trustees in the original position from that of free and equal citizens in a well-ordered society, who may well be morally motivated by their sense of justice to do what is right and just for its own sake.)

Three factors then play a role in motivating the parties in the original position: (1) First, they aim to advance their determinate conception of the good, or rational plan of life, even though they do not know what that conception is. Moreover, they also seek conditions that enable them to exercise and develop their “moral powers,” namely (2) their rational capacities to form, revise and pursue a conception of their good, and (3) their capacity to be reasonable and to have a sense of justice. These are the three “higher-order interests” the parties to Rawls's original position aim to promote in their agreement on principles of justice.

The three higher-order interests provide the basis for Rawls's account of *primary social goods*. The primary goods are the all-purpose social means that are necessary to the exercise and development of the moral powers and to pursue a wide variety of conceptions of the good. Rawls describes them initially in *Theory* as goods that any rational person should want, whatever his or her rational plan of life. The primary social goods are: rights and liberties, powers and opportunities, income and wealth, and the social bases of self-respect. ‘Powers’ refer not simply to a capacity to effect outcomes or influence others' behavior. Rawls rather uses the term to refer to the legal and institutional abilities and prerogatives that attend offices and social positions. Hence, he sometimes refers to the primary goods of “powers and prerogatives of offices and positions of authority and responsibility” (JF 58). Members of various professions and trades have institutional powers and prerogatives that are characteristic of their position and which are necessary if they are to carry out their respective roles and responsibilities. By income and wealth Rawls says he intends “all-purpose means” having an exchange value, which are generally needed to achieve a wide range of ends (JF 58–59). Finally, “the social bases of self-respect” are features of institutions that are needed to enable people to have the confidence that they and their position in society are respected and that their conception of the good is worth pursuing and achievable. These features depend upon history and culture. Primary among these social bases in a democratic society are the conditions needed for equal citizenship, including equality of political rights and fair equal opportunity, as well as personal independence and adequate material means for achieving it. The social bases of self-respect are crucial to Rawls's argument for *equal* basic liberties, especially political equality and equal rights of political participation.

The parties to the original position are motivated to achieve a fully adequate share of primary goods so they can achieve their higher-order interests in pursuing their rational plans of life and exercising their moral powers. “They assume that they normally prefer more primary social goods rather than less” (TJ, 142/123
This too is part of being rational. Because they are not envious, their concern is with the absolute level of primary goods, not their share relative to other persons.

To sum up, the parties in the original position are formally rational in that they are assumed to have and to effectively pursue a rational plan of life with a schedule of coherent purposes and commitments that give their life meaning. As part of their rational plans, they have a substantive interest in the development and exercise of their capacities to be rational and to be reasonable. These “higher-order interests” provide them with reason to procure for themselves in their choice of principles of justice an adequate share of the primary social goods that enable them to achieve these higher-order ends and effectively pursue their life plans.

A final feature of Rawls's account of rationality is a normal human tendency he calls “the Aristotelian principle” (TJ sect.65). This “deep psychological fact” says that, other things equal, people normally find activities that call upon the exercise of their developed capacities to be more interesting and preferable to engaging in simpler tasks, and their enjoyment increases the more the capacity is developed and realized or the greater its complexity (TJ, 426/374). Moreover Rawls believes that development of our “higher capacities” (J.S. Mill's term) is important to our sense of self-respect. These general facts imply that rational people should incorporate into their life plans activities that call upon the exercise and development of their talents and skills and distinctly human capacities (TJ 432/379). This motivation is especially relevant to Rawls's argument for the stability of justice as fairness and the good of justice. (See below, Section 5.3.) The important point here is that the Aristotelian principle is taken into account by the parties in their decision on principles of justice. They want to choose principles that maintain their self-respect and enable them to freely develop their human capacities and pursue a wide range of activities, as well as engage their capacities for a sense of justice.

5. Other Conditions on Choice in the Original Position

The veil of ignorance is the primary condition that constrains the rational choice of the parties in the original position. There are several other conditions imposed on their agreement.

5.1 The Circumstances of Justice (TJ §22)

Among the general facts the parties know are “the circumstances of justice.” Rawls says these are “conditions under which human cooperation is both possible and necessary” (TJ 126/109 rev.). Following Hume, Rawls distinguishes two general kinds: the objective and subjective circumstances of justice. The former include physical facts about human beings, such as their rough similarity in mental and physical faculties, and vulnerability to the united force of others. Objective circumstances also include conditions of moderate scarcity of resources: there are not enough resources to satisfy everyone's demands, but there are enough to provide all with adequate satisfaction of their basic needs; unlike conditions of extreme scarcity (e.g. famine) cooperation then seems productive and worthwhile for people.

Among the subjective circumstances of justice is the parties' mutual disinterestedness, which reflects the “limited altruism” (TJ 146/127) of persons in society. Free and equal persons have their own plans of life and special commitments to others, as well as different philosophical and religious beliefs and moral doctrines (TJ 127/110). Hume says that if humans were impartially benevolent, equally concerned with everyone's welfare, then justice would be unnecessary. People then would willingly sacrifice their interests for the greater advantage of other. They would not be concerned about their personal rights or possessions, and property would be unnecessary (Hume, 1777, 185–186). But we are more concerned with our own aims and interests—which include our interests in the interests of those nearer and dearer to us—than we are with the interests of strangers with whom we have few if any interactions. This implies a potential conflict of human interests. Rawls adds that concern for our interests and plans of life does not mean we are selfish or have
interests only in ourselves—interests of a self should not be confused with interests in oneself; we have interests in others and in all kinds of causes and ends (TJ 127/110). But, as history shows, our benevolent interests in others and in religious and philosophical doctrines are as often the cause of conflict as is self-interest.

The subjective circumstances of justice also include limitations on human knowledge, thought, and judgment, as well as emotional influences and great diversity of experiences. These lead to biases and inevitable disagreements in factual and other judgments, as well as to differences in religious, philosophical, and moral convictions. In *Political Liberalism*, Rawls highlights these subjective circumstances, calling them “the burdens of judgment.” They imply, significantly, that regardless how impartial and altruistic people are, they still will disagree in their factual judgments and in religious, philosophical, and moral doctrines. Disagreements in these matters are inevitable even among rational and reasonable people. This is “the fact of reasonable pluralism,” which is another general fact known to the parties in the original position. Reasonable pluralism of doctrines lends significant support to Rawls's arguments for the first principle of justice, especially to equal basic liberties of conscience, expression, and association.

### 5.2 Publicity and other Formal Constraints of Right (TJ §23)

There are five “formal constraints” associated with the concept of right that Rawls says the parties must take into account in coming to agreement. The more a conception of justice satisfies these formal constraints of right, the more reason the parties have to choose that conception. The formal constraints of right are: generality, universality in application, ordering of conflicting claims, publicity, and finality. The *ordering condition* says that a conception of justice should aspire to completeness: it should be able to resolve conflicting claims and order their priority. Ordering implies a systematicity requirement: principles of justice should provide a determinate resolution to problems of justice that arise under them; and in so far as a conception of justice is not able to order conflicting claims and resolve problems of justice, that is a reason against choosing it in the original position. The ordering condition is important in Rawls's argument against pluralist moral doctrines he calls “Intuitionism.”

The *publicity condition* says that the parties are to assume that the principles of justice they choose will be publicly known and recognized as the basis for social cooperation among the people whose relations they organize and regulate. This implies that people will not be uninformed or have false beliefs about the bases of their social and political relations. There are to be no “noble lies” or false ideologies obscuring a society's principles of justice. Publicity of principles of justice is required to respect persons as free and equal citizens. Rawls believes that people should know the bases of their social and political relations, and not have to be deceived about them in order to cooperate and live together on fair terms. This condition plays an important role in Rawls's arguments against utilitarianism and other consequentialist conceptions.

Related to publicity is that principles should be *universal in application*. This implies not simply that “they hold for everyone in virtue of their being moral persons” (TJ 132/114 rev.). It also means that everyone can understand the principles of justice and use them in their deliberations. Universality in application then imposes a limit on how complex principles of justice can be—they must be understandable to common moral sense, and not so complicated that only experts can apply them in deliberations. For among other things, these principles are to guide democratic citizens in their judgments and shared deliberations about just laws and policies.

Both publicity and universality in application (as Rawls defines it) are controversial conditions. Utilitarians, for example, have argued that the truth about morality and justice is so complicated and controversial that it might be necessary to keep it hidden from most individuals' awareness. For morality often requires much that is contrary to their personal interests. Also sometimes it's just too complicated for people to understand why their moral duties require of them what they do. So long as they understand their individual duties, it may be necessary to keep the full truth hidden from them. This, however, is not necessary in Rawls's way of thinking.
better if they do not understand the principles and reasons behind them. So Sidgwick argues that the aims of utilitarianism might better be achieved if it remains an “esoteric morality,” knowledge of which is confined to “an enlightened few” (Sidgwick, 489–90). The reason Rawls sees publicity and universality as necessary relates to the conception of the person implicit in justice as fairness. If we conceive of persons as free and equal moral persons capable of rational and moral autonomy, then they should not be under any illusions about the bases of their social relations, but should be able to understand and apply these principles in their deliberations about justice. These are important conditions of the freedom and autonomy (moral and political) of democratic citizens.

5.3 The Stability Requirement

Rawls says, “An important feature of a conception of justice is that it should generate its own support. Its principles should be such that when they are embodied in the basic structure of society, people tend to acquire the corresponding sense of justice and develop a desire to act in accordance with its principles. In this case a conception of justice is stable” (TJ, 138/119). The parties in the original position are to take into account the “relative stability” of a conception of justice and the society that institutes it. The stability of a just society does not mean that it must be unchanging. It means rather that in the face of inevitable change members of a society should be able to maintain their allegiance to principles of justice and the institutions they support. When disruptions to society do occur (via economic crises, war, natural catastrophes, etc.) and/or society departs from justice, citizens' commitments to principles of justice are sufficiently robust that just institutions are eventually restored. The role of the stability requirement for Rawls is twofold: first, to test whether potential principles of justice are compatible with human natural propensities, or our moral psychology and general facts about social and economic institutions; and second, to determine whether principles are conducive to realizing the human good. To be stable principles of justice should be realizable in a feasible and enduring social world. They need to be practicably possible given the limitations of the human condition. Moreover, this feasible social world must be one that can endure over time, not by just any means, but by gaining the willing support of people who live in it. People should knowingly want to uphold and maintain society's just institutions not just because they benefit from them, but on grounds of their sense of justice. In choosing principles of justice, the parties in the original position must take into account their “relative stability.” They have to consider the degree to which a conception (in comparison with other conceptions) describes an achievable and sustainable system of social cooperation, and whether the institutions and demands of such a society will attract people's willing compliance and generally engage their sense of justice.

For example, suppose principles of justice were to impose a duty to practice impartial benevolence towards all people, and thus a duty to show no greater concern for the welfare of ourselves and loved ones than we do towards millions if not billions of others. This principle demands too much of human nature and would not be feasible—people simply would reject its onerous demands. But Rawls's stability requirement implies more than just ‘ought implies can.’ It says that principles of justice and the scheme of social cooperation they describe should evince “stability for the right reasons” (PL xliii, CP 589). Recall here the higher-order interests of the parties in development and exercise of their capacities for justice. A just society should be able to endure not simply as a modus vivendi, by coercive enforcement of its provisions and its promoting the majority of peoples' interests. Stability “for the right reasons” requires that people support society for moral reasons of justice; society's basic principles must respond to reasonable persons' capacities for justice and engage their sense of justice. Rawls regards our moral capacities for justice as an integral part of our nature as sociable beings. He believes that one role of a conception of justice is to accommodate human capacities for sociability, the capacities for justice that enable us to be cooperative social beings. So not only should a conception of justice advance human interests, but it should also answer to our moral psychology by enabling us to knowingly and willingly exercise our moral capacities and sensibilities, which are among the moral
powers to be reasonable. This is one way that Rawls's conception of justice is "ideal-based" (CP 400-401 n.): it is based in an ideal of human beings as free and equal moral persons and an ideal of their social relations as acceptable and justifiable to them (the ideal of a well-ordered society).

This relates to the second ground for the stability condition, which can only be mentioned here: it is that principles of justice should be compatible with, and even conducive to, the human good. It speaks strongly in favor of a conception of justice that it is compatible with and promotes the human good. First, if a conception of justice requires of many reasonable people that they change their conscientious philosophical or religious convictions for the sake of satisfying a majority's beliefs, or abandon their pursuit of the important interests that constitute their plan of life, this conception could not gain their support and would not be stable over sustained periods of time. Moreover, Rawls assumes that a conception of justice should enable citizens to adequately exercise and fully develop their moral powers. It must then engage their sense of justice; ideally, they should not regard justice as a burden but should come to experience that acting on and from principles of justice is worth doing for its own sake. For Rawls, it speaks strongly in favor of a conception of justice that acting for the sake of its principles is experienced as an activity that is good in itself. For then justice and exercise of the sense of justice are for those persons intrinsic goods and a precondition for their living a good life.

Now turn to the arguments for the principles of justice in the original position.

6. The Arguments for the Principles of Justice from the Original Position

The original position is not a bargaining situation where the parties make proposals and counterproposals and negotiate over different principles of justice. Nor is it a wide ranging discussion where the parties deliberate and design their own conception of justice (unlike, for example, Habermas' discourse ethics; see Habermas, 1995). Instead, the parties' deliberations are much more constrained. They are presented with a list of conceptions of justice taken from the tradition of western political philosophy. These include different versions of utilitarianism, perfectionism, and intuitionism (or pluralist views), rational egoism, justice as fairness, and a group of "mixed conceptions" that combine elements of these (For Rawls's initial list see TJ 124/107) (Rawls later says libertarianism should also be added to the list, and the principles of justice would still be agreed to., JF 83; Nozick,198ff., agrees and says the OP is incapable of yielding entitlement principles due to the self-interest of the parties.). The parties' deliberations are confined to discussing and agreeing upon the conception that each finds most rational, given their specified interests. In a series of pairwise comparisons, they consider all the conceptions of justice made available to them and ultimately agree unanimously to accept the conception that survives this winnowing process. In this regard, the original position is best seen as a kind of selection process wherein the parties' deliberations are constrained by the background conditions imposed by the original position as well as the list of conceptions of justice provided to them. They are assigned the task of agreeing on principles for designing the basic structure of a self-contained society under the circumstances of justice.

In making their decision, the parties are motivated only by their own rational interests. They do not take moral considerations of justice into account except in so far as these considerations bear on their achieving their interests. Their interests again are defined in terms of their each acquiring an adequate share of primary social goods (rights and liberties, powers and opportunities, income and wealth, etc.) and achieving the background social conditions enabling them to effectively pursue their conception of the good and realize their higher-order interests in developing and exercising their moral powers. Since the parties are ignorant of their particular conceptions of the good and of all other particular facts about their society, they are not in a position to engage in bargaining. In effect they all have the same general information and are motivated by
Rawls makes four arguments in *Theory*, Part I for the principles of justice. The main argument for the difference principle is made later in section 49, and is substantially amended and clarified in *Justice as Fairness: A Restatement*. The common theme throughout the original position arguments is that it is more rational for the parties to choose the principles of justice over any other alternative. Rawls devotes most of his attention to the comparison of justice as fairness with classical and average utilitarianism, with briefer discussions of perfectionism (TJ, sect. 50) and intuitionism. Here I'll focus discussion on Rawls's comparison between justice as fairness and utilitarianism.

6.1 The Argument from the Maximin Criterion (TJ sects. 26–28)

Describing the parties' choice as a rational choice subject to the reasonable constraints imposed by the original position allows Rawls to invoke the theory of rational choice and decision under conditions of uncertainty. In rational choice theory there are a number of potential “strategies” or rules of choice that are more or less reliably used depending on the circumstances. One rule of choice—called “maximin”—directs that we play it as safe as possible by choosing the alternative whose worst outcome leaves us better off than the worst outcome of all other alternatives. The aim is to “maximize the minimum” regret or loss to well-being. To follow this strategy, Rawls says you should choose as if your enemy were to assign your social position in whatever kind of society you end up in. By contrast another strategy leads us to focus on the most advantaged position and says we should “maximize the maximum” potential gain—“maximax”—and choose the alternative whose best outcome leaves us better off than all other alternatives. Which, if either, of these strategies is more sensible to use depends on the circumstances and many other factors.

A third strategy advocated by orthodox Bayesian decision theory, says we should always choose to directly maximize expected utility. To do so under conditions of uncertainty of outcomes, the degree of uncertainty should be factored into one's utility function, with probability estimates assigned to alternatives based on the limited knowledge that one has. Given these subjective estimates of probability incorporated into one's utility function, one can always choose the alternative that maximizes expected utility. Since it simplifies matters to apply the same rule of choice to all decisions this is a highly attractive idea, so long as one can accept that it is always safe to assume that the maximization of expected utility leads over time to maximizing actual utility.

What about those extremely rare instances where there is absolutely no basis upon which to make probability estimates? Suppose you don't even have a hunch regarding the greater likelihood of one alternative over another. According to orthodox Bayesian decision theory, the “principle of insufficient reason” then should be observed; it says that when there is no reason to assign a greater likelihood to one alternative rather than another, then an equal probability is to be assigned to each potential outcome. This makes sense on the assumption that if you have no more premonition of the likelihood of one option rather than another, they are for all you know equally likely to occur. By observing this rule of choice consistently over time, a rational chooser presumably should maximize his or her individual expected utility, and hopefully actual utility as well.

What now is the appropriate decision rule to be used to choose principles of justice under conditions of complete uncertainty of probabilities in Rawls's original position? Rawls argues that, given the enormous gravity of choice in the original position, plus the fact that the choice is not repeatable, it is rational for the parties to follow the maximin strategy when choosing between the principles of justice and principles of average or aggregate utility (or most any other purely consequentialist principle). Not surprisingly, following the maximin rule of choice results in choice of the principles of justice over the principles of utility (average or aggregate); for unlike utilitarianism, justice as fairness guarantees equal basic liberties, fair equal opportunities, and an adequate social minimum for all citizens.
Why does Rawls think maximin is the rational choice rule? Recall what is at stake in choice from the original position. The decision is not an ordinary choice. It is rather a unique and irrevocable choice where the parties decide the basic structure of their society, or the kind of social world they should live in and the background conditions against which they will develop and pursue their aims. It is a kind of superchoice—an inimitable choice of the background conditions for all one's future choices. Rawls argues that because of the unique importance of the choice in the original position—including the gravity of the choice, the fact that it is not renegotiatable or repeatable, and the fact that it determines all one's future prospects—it is rational to follow the maximin rule and choose the principles of justice. For should even the worst transpire, the principles of justice provide an adequate share of primary goods enabling one to maintain one's conscientious convictions and sincerest affections and pursue a wide range of permissible ends by protecting equal basic liberties and fair equal opportunities and guaranteeing a social minimum of income and wealth. The principles of utility, by contrast, provide no guarantee of any of these goods.

Rawls says that in general there are three conditions that must be met in order to make it rational to follow the maximin rule (TJ 154–55/134). First, there should be no basis or at most a very insecure basis upon which to make estimates of probabilities. Second, the choice singled out by observing the maximin rule is an acceptable alternative we can live with, so that one cares relatively little by comparison for what is to be gained above the minimum conditions secured by the maximin choice. When this condition is satisfied, then no matter what position one eventually ends up in, it is at least acceptable. The third condition for applying the maximin rule is that all the other alternatives have (worse) outcomes that we could not accept and live with. Of these three conditions Rawls later says that the first plays a minor role, and that it is the second and third conditions that are crucial to the maximin argument for justice as fairness (JF 99). This seems to suggest that, even if the veil of ignorance were not as thick and parties did have some degree of knowledge of the likelihood of ending up in one social position rather than another, still it would be more rational to choose the principles of justice over the principle of utility.

Rawls contends all three conditions for the maximin strategy are satisfied in the original position when choice is made between the principles of justice and the principle of utility (average and aggregate). Because all one's values, commitments, and future prospects are at stake in the original position, and there is no hope of renegotiating the outcome, a rational person would agree to the principles of justice instead of the principle of utility. For the principles of justice imply that no matter what position you occupy in society, you will have the rights and resources needed to maintain your valued commitments and purposes, to effectively exercise your capacities for rational and moral deliberation and action, and to maintain your sense of self-respect as an equal citizen. With the principle of utility there is no such guarantee; everything is “up for grabs” (so to speak) and subject to loss if required by the greater sum of utilities. Conditions (2) and (3) for applying maximin are then satisfied in the comparison of justice as fairness with the principle of (average or aggregate) utility.

It is often claimed that Rawls's parties are “risk-averse;” otherwise they would never follow the maximin rule but would take a chance on riskier but more rewarding outcomes provided by the principle of utility. Thus, John Harsanyi contends that it is more rational under conditions of complete uncertainty always to choose according to the principle of insufficient reason and assume an equal probability of occupying any position in society. When the equiprobability assumption is made, the parties in the original position would choose the principle of average utility instead of the principles of justice (Harsanyi 1975).

Rawls denies that the parties have a psychological disposition to risk-aversion. He argues however that it is rational to choose as if one were risk averse under the highly exceptional circumstances of the original position. His point is that, while there is nothing rational about a fixed disposition to risk aversion, it is nonetheless rational in some circumstances to choose conservatively to protect certain fundamental interests against loss or compromise. It does not make one a risk averse person, but instead normally it is entirely rational to purchase auto liability, health, home, and life insurance against accident or calamity. The original
position is such a situation writ large. Even if one knew in the original position that the citizen one represents enjoys taking risks, this would still not be a reason to gamble with his or her rights, liberties and starting position in society. For if the risktaker were born into a traditional, repressive, or fundamentalist society, she might well have little opportunity for taking the kinds of risks, such as gambling, that she normally enjoys. It is rational then even for risktakers to choose conservatively in the original position and guarantee their future opportunities to gamble or otherwise take risks.

Harsanyi and other orthodox Bayesians contend that maximin is an irrational decision rule, and provide ample examples. But examples do not suffice here; simply because maximin is under many circumstances irrational does not mean that it is never rational. No doubt maximin is an irrational strategy under most circumstances of choice uncertainty, particularly under circumstances where we will have future opportunities to recoup our potential losses and choose again. But these are not the circumstances of the original position; once the rules of justice are decided, they apply in perpetuity, and there is no opportunity to renegotiate or escape the situation. One who relies on the equiprobability assumption in choosing principles of justice in the original position is being foolishly reckless given the gravity of choice at stake. It is not being risk-averse, but rather entirely rational to be unwilling to gamble with one's basic liberties, fair opportunities and adequate resources needed to pursue one's most cherished ends and commitments, simply for the sake of gaining the marginally greater social powers, income and wealth that might be available to some in a society governed entirely by the principle of utility.

Rawls exhibits the force of the maximin argument in discussing liberty of conscience. He says (TJ, sect. 33) that a person who is willing to jeopardize the right to hold and practice his or her conscientious religious, philosophical and moral convictions, all for the sake of gaining uncertain added benefits via the principle of utility, does not know what it means to have conscientious beliefs, or at least does not take such beliefs seriously (TJ 207–08/181–82). A rational person with convictions about what gives life meaning is not willing to negotiate with and gamble away the right to hold and express those convictions and the freedom to act on them. After all what could be the basis for negotiation, for what could matter more? Some people (e.g. some nihilists) may not have any conscientious convictions, and are simply willing to act on impulse or on whatever thoughts and desires they happen to have at the moment. But behind the veil of ignorance no one knows whether he or she is such a person, and there are no grounds for making this assumption. The parties must take into account that they might have convictions and values they are unwilling to compromise. (Besides, the nihilist should want to protect his freedom to be the person he is; what if he chose the principle of utility and ended up in a religiously fundamentalist society governed by Puritans?) Thus it remains irrational to jeopardize basic liberties by choosing the principle of utility instead of the principles of justice.

None of this is to say that maximin is normally a rational choice strategy. Rawls himself says it “is not, in general, a suitable guide for choices under uncertainty” (TJ 153). As we see below in Section 6.4, it is not even a rational strategy in the original position when the alternatives for choice, unlike the principle of utility, guarantee basic liberties and a social minimum. Rawls relies upon the maximin argument mainly to argue for the first principle of justice and a guaranteed social minimum. Other arguments are needed to justify the difference principle.

### 6.2 The Strains of Commitment

There are three additional arguments Rawls makes to support justice as fairness (all in TJ, sect. 29). Each of these depends on the concept of a “well-ordered society.” There are two sides to Rawls's social contract. The parties in the original position have the task of agreeing to principles that all rationally can accept under the circumstances of the original position. But their rational choice is partially determined by the principles that free and equal moral persons reasonably can accept and agree to in a well-ordered society. The ideal of a well-ordered society is Rawls's development of social contract doctrine. It is a society in which (1) everyone
willingly accepts and agrees to the same principles of justice; (2) these principles are successfully realized in basic social institutions and generally are complied with; and (3) reasonable persons are morally motivated to comply by their sense of justice (TJ 4–5, §69).

The first of Rawls's three arguments highlights the idea that choice in the original position is an agreement, and involves certain "strains of commitment," assumed by all to comply with the principles they agree to (TJ 176f./153f. and CP 250ff). Knowing that they are required to choose principles for a well-ordered society, the parties must choose principles that they sincerely believe they will be able to accept, endorse and willingly comply with under conditions where these principles are generally enforced. For reasons to be discussed shortly, Rawls says this condition favors agreement on the principles of justice over utilitarianism and other alternatives.

But first, consider the frequent objection that there is no genuine agreement at all in the original position, for the thick veil of ignorance deprives the parties of all bases for bargaining (cf. TJ, 139–40/120–21 rev.). In the absence of bargaining, it is said, there can be no contract. For contracts must involve a quid pro quo—something given for something received (called ‘consideration’ at common law). The parties in the OP cannot bargain without knowing what they have to offer or to gain in exchange. So (the objection continues) Rawls's original position does not involve a real contract; rather, since the parties are all “described in the same way,” it is simply the rational choice of one person. (See Hampton, 1980, 334; see also Gauthier, 1985, 203.)

In response, not all contracts involve bargaining or are of the nature of economic transactions. Some involve a mutual pledge and commitment to shared purposes and principles. Marriage contracts, or agreements among friends or the members of a religious, benevolent, or political association are often of this nature. For example, the Mayflower Compact was a “covenant” to “combine ourselves together into a civil body politic” charged with making and administering “just and equal laws…for the general good.” The agreement in Rawls's original position is more of this nature. Even though ignorant of particular facts about themselves, the parties in fact do give something in exchange for something received: they all exchange their mutual commitment to accept and abide by the principles of justice and to uphold just institutions once they enter their well-ordered society. Each agrees only on condition others do too, and all tie themselves into social and political relations in perpetuity. Their agreement is final, and they will not permit its renegotiation should circumstances turn out to be different than some hoped for. Their mutual commitment to justice is reflected by the fact that once these principles become embodied in institutions there are no legal means that permit anyone to depart from the terms of their agreement. As a result, the parties have to take seriously the legal obligations and social sanctions they will incur as a result of their agreement, for there is no going back to the initial situation. So if they do not sincerely believe that they can accept the requirements of a conception of justice and conform their actions and life plans accordingly, then these are strong reasons to avoid choosing those principles. It would not be rational for a party to take risks, falsely assuming that if he ends up badly, he can violate at will the terms of agreement or later regain his initial situation and renegotiate terms of cooperation. (See Freeman, 1990; Freeman, 2007b, 180–182.)

Rawls gives special poignancy to this mutual commitment of the parties by making it a condition that the parties cannot choose and agree to principles in bad faith; they have to be able, not simply to live with and grudgingly accept, but instead to willingly endorse the principles of justice as members of society. Essential to Rawls's argument for stability is the assumption of everyone's willing compliance with requirements of justice. This is a feature of a well-ordered society. The parties are assumed to have a sense of justice; indeed the development and exercise of it is one of their fundamental interests. Hence they must choose principles that they can not only accept and live with, but which are responsive to their sense of justice and they can unreservedly endorse. Given these conditions on choice, a party cannot take risks with principles he knows he will have difficulty complying with voluntarily. He would be making an agreement in bad faith, and this is ruled out by the conditions of the original position.
Rawls contends that these “strains of commitment” created by the parties' agreement strongly favor the principles of justice over the principles of utility and other teleological views. For everyone's freedom and basic needs are met by the principles of justice because of their egalitarian nature. Given the lack of these guarantees under the principle of utility, it is much more difficult for those who end up worse off in a utilitarian society to willingly accept their situation and commit themselves to the utility principle. It is a rare person indeed who can freely and without resentment sacrifice his or her life prospects so that those who are better off can have even greater comforts, privileges, and powers. This is too much to demand of our capacities for human benevolence. It requires a kind of commitment that people cannot make in good faith, for who could willingly support laws that are so detrimental to oneself and the people one cares about most that they must sacrifice their fundamental interests for the sake of those more advantaged? Besides, why should we encourage such subservient dispositions and the accompanying lack of self-respect? The principles of justice, by contrast, conform better with everyone's interests, their desire for self-respect and their natural moral capabilities to reciprocally recognize and respect others' legitimate interests while freely promoting their own good. The strains of commitment incurred by agreement in the original position provide strong reasons for the parties to choose the principles of justice and reject the risks involved in choosing the principles of average or aggregate utility.

6.3 Stability, Publicity, and Self-Respect

Rawls's strains-of-commitment argument explicitly relies upon a rarely noted feature of his argument: it involves in effect two social contracts. First, hypothetical agents situated equally in the original position unanimously agree to principles of justice. This agreement has attracted the most attention from Rawls's critics. But the parties' hypothetical agreement in the original position is patterned on the general acceptability of a conception of justice by free and equal persons in a well-ordered society. Rawls says, “The reason for invoking the concept of a contract in the original position lies in its correspondence with the features of a well-ordered society [which] require…that everyone accepts, and knows that the others accept, the same principles of justice” (CP 250). In order for the hypothetical parties in the original position to agree on principles of justice, there must be a high likelihood that real persons, given human nature and general facts about social and economic cooperation, can also agree and act on the same principles, and that a society structured by these principles is feasible and can endure. This is the stability requirement referred to earlier. One conception of justice is relatively more stable than another the more willing people are to observe its requirements under conditions of a well-ordered society. Assuming that each conception of justice has a corresponding society that is as well-ordered as can be according to its terms, the stability question raised in Theory is: Which conception of justice is more likely to engage the moral sensibilities and sense of justice of free and equal persons as well as affirm their good? This requires an inquiry into moral psychology and the human good, which takes up most of Part III of A Theory of Justice.

Rawls makes two arguments from the original position that invoke the stability requirement., the arguments (1) from publicity and (2) from self-respect.

(1) The argument from publicity: Rawls contends that utilitarianism, perfectionism, and other “teleological” conceptions are not likely to be freely acceptable to many citizens when made fully public under the conditions of a well-ordered society. Recall the publicity condition discussed earlier: A feature of a well-ordered society is that its regulative principles of justice are publicly known and appealed to as a basis for deciding laws and justifying basic institutions. A conception of justice that cannot satisfy this condition is to be rejected by the parties. Rawls contends that under the publicity condition justice as fairness generally engages citizens' sense of justice and remains more stable than utilitarianism (TJ 177f./154f.). For public knowledge that reasons of maximum average (or aggregate) utility determine the distribution of benefits and burdens would lead those worse-off to object to and resent their situation. After all, their well-being and interests, perhaps even their basic liberties, are being sacrificed for the greater happiness of those who are
already more fortunate and have a greater share of primary social goods. It is too much to expect of human nature that people should freely acquiesce in and embrace such terms of cooperation. By contrast, the principles of justice are designed to advance reciprocally everyone's position; those who are better off do not achieve their gains at the expense of the less advantaged. “Since everyone's good is affirmed, all acquire inclinations to uphold the scheme” (TJ, 177/155). It is a feature of our moral psychology, Rawls contends, that we normally come to form attachments to people and institutions that are concerned with our good; moreover we tend to resent those persons and institutions that act contrary to our good. Rawls argues at length in chapter 8 of Theory, §§70–75, that justice as fairness accords with the reciprocity principles of moral psychology that are characteristic of human beings' moral development.

(2) The argument from the social bases of self-respect: The publicity condition is also crucial to Rawls' fourth argument for the principles of justice, from the social bases of self-respect (TJ, 178–82/155–59 rev.). These principles, when publicly known, give greater support to citizens' sense of self-respect than do utilitarian and perfectionist principles. Rawls says self-respect is “perhaps the most important primary good,” (TJ, 440/386 rev.) since few things seem worth doing if a person has little sense of his or her own worth or no confidence in his or her abilities to execute a worthwhile life plan. The parties in the original position will then aim to choose principles that best secure the sense of self-respect. Now being regarded by others as a free and independent person of equal status with others is crucial to the self-respect of persons who regard themselves as free and equal. Justice as fairness, by affording and protecting the priority of equal basic liberties and fair equal opportunities for all, secures the status of each as free and equal citizens. For example, because of equal political liberties, there are no “passive citizens” who must depend on others to politically protect their rights and interests; and with fair equal opportunities no one has grounds to experience the resentment that inevitably arises in societies where social positions are effectively closed to those less advantaged or less powerful. Moreover, the second principle secures adequate social powers and resources for all to make everyone's equal basic liberties worthwhile. It has the effect of making citizens socially and economically independent, so that no one need be subservient to the will of another. Citizens then can regard and respect one another as equals, and not as masters or subordinates. ("Non-domination," an idea central to contemporary Republicanism, is then essential to citizens' sense of self-respect in Rawls's sense. See Pettit, 1997) Equal basic liberties, fair equal opportunities, and political and economic independence are primary among the social bases of self-respect in a democratic society. The parties in the original position should then choose the principles of justice over utilitarianism and other teleological views both to secure their sense of self-respect, and to procure the same for others, thereby guaranteeing greater overall stability.

Rawls substantially relies on the publicity condition to argue against utilitarianism and perfectionism. He says publicity “arises naturally from a contractarian standpoint” (TJ, 133/115 rev.). In Theory he puts great weight on publicity ultimately because he thinks that giving people knowledge of the moral bases of coercive laws is a condition of fully acknowledging and respecting them as free and responsible rational moral agents. With publicity of first principles, people have knowledge of the real reasons for their social and political relations and the formative influences of the basic structure on their characters, plans and prospects. In a well-ordered society with a public conception of justice, there is no need for an “esoteric morality” that must be confined “to an enlightened few” (as Sidgwick says of utilitarianism, Sidgwick, 490). Moreover, public principles of justice can serve agents in their practical reasoning and provide democratic citizens a common basis for political argument and justification. These considerations underlie Rawls's later contention that having knowledge of the principles that determine the bases of social relations is a precondition of individuals' freedom. (CP 325f.) Rawls means that publicity of society's fundamental principles is a condition of citizens' exercise of the powers and abilities that enable them to take full responsibility for their lives. Full publicity is then a condition of the political, rational, and moral autonomy of persons, which are all significant values according to justice as fairness.

6.4 The Argument for the Difference Principle
Rawls's second principle of justice says:

Social and economic inequalities are to be arranged so that they are both: (a) to the greatest benefit of the least advantaged...and (b) attached to offices and positions open to all under conditions of fair equality of opportunity. (TJ 302/266)

The difference principle (the first part of the second principle, through (a)), addresses differences or inequalities in the distribution of the primary goods of income and wealth and powers and positions of office and responsibility. It basically requires that a society is to institute the economic system that would make the least advantaged class better off than they would be in any other feasible economic system, compatible with maintaining citizens equal basic liberties and fair equality of opportunity. Rawls defines “least advantaged” as those with the least share of the primary goods of income and wealth, and powers and positions of office. He also assumes the least advantaged are “fully cooperating” (JF 179), and “full and active participants in society” (CP, 259) who do their fair share in contributing to economic activity and the social product. Rawls thus regards “distributive shares” under the difference principle as the benefits that accrue to persons for doing their part in socially productive cooperation. The class of unskilled workers receiving minimum income satisfies his definition of least advantaged (TJ 98/84 rev.). Those who are able but unwilling to work “must somehow support themselves” (JF 179; cf. PL 182n.). Disability payments to meet the needs of severely handicapped citizens who are permanently unable to work are not ascertained by reference to the difference principle for Rawls, but instead are to be determined by principles of assistance, which he leaves unspecified (PL 272n.). As a principle for structuring basic economic institutions on grounds of reciprocity, the difference principle itself is not an appropriate principle for compensating handicaps and other disabilities.

Rawls contends it is an empirical question which economic system satisfies the difference principle. He claims however that under ideal conditions of a well-ordered society, where the principles of justice are generally accepted, either a “property-owning democracy” or liberal (market) socialism are most likely to satisfy the difference principle, to the exclusion of laissez-faire capitalism, command economy communism, and even the modern capitalist welfare state. Property-owning democracy (POD) differs from the capitalist welfare state mainly in that it is not marked by such broad discrepancies in income and wealth, but instead provides for widespread ownership and control of productive wealth and means of production, including opportunities for greater freedom and control in the workplace. Thus, akin to market socialism, POD is not marked by sharp divisions between capital and labor but manifests broad distribution of economic powers and positions, as well as a more equal distribution of income and wealth. (See JF 135–140, 158–162, 178.) For these reasons, and since a property-owning democracy also secures the fair value of equal political liberties and fair equality of opportunities, a property-owning democracy provides a more secure basis for citizens' sense of self-respect (the remaining primary social good covered by the difference principle in part) than does welfare-state capitalism. (See Freeman, 2007b, 219–235.)

Rawls relies on the maximin rule of choice to argue against the principle of utility. Since the maximin rule and the difference principle both require maximizing the minimum position, it seems natural to assume that the maximin choice rule leads directly to choice of the difference principle in the original position. Though Rawls might have conveyed this impression in Theory (§26), he later says that in fact the maximin rule alone cannot be used to justify the difference principle (JF, 43n). For when the difference principle is compared with those “mixed conceptions” of economic justice that provide some form of equal opportunities and a social minimum, then the third condition for applying maximin is not satisfied (see above, sect. VI,A). The third condition for the maximin rule implies that there can only be one acceptable alternative for choice. If there is a second alternative whose least advantaged position is one that rational persons can live with and accept if they end up in the least advantaged position, then the maximin rule is not a rational rule of decision. For then there is no likelihood of grave risks to one's future prospects.
“Mixed conceptions” of justice (see TJ §49) protect the basic liberties, and in most cases provide equal opportunities and guarantee a social minimum of income and wealth sufficient to meet basic needs. But they determine the social minimum in some way other than the difference principle. For example, the social minimum may be determined by balancing citizens' moral intuitions regarding a decent minimum; or according the rule that the adequate social minimum is one-third of the median market income in society; or that it is whatever amount that is needed to maintain social stability and avoid disruption by the less advantaged. Once the social minimum is satisfied mixed conceptions might then rely upon the principle of average utility (or some alternative) to decide economic policies and distributions. Rawls calls this the “principle of restricted utility,” (JF 120, 126) since the pursuit of social utility is restricted by the basic liberties, fair opportunities, and a fixed social minimum. This is one of several possible mixed conceptions that combine the first principle of justice with a principle of distributive justice with a social minimum other than the difference principle (TJ, 124/107 rev.). Rawls apparently regards the capitalist welfare state as based in mixed conceptions that incorporate such a principle of restricted utility (JF 120–130, 139–140).

Rawls concedes that “mixed conceptions are much more difficult to argue against than the principle of utility,” since “the strong arguments from liberty cannot be used as before” (TJ, 316/278 rev.). He discusses mixed conceptions in Theory, §49, and devotes more attention to them later in Justice as Fairness: A Restatement (§§34–38.). Rawls makes one main argument in favor of the difference principle and several more specific arguments against the principle of restricted utility. The main argument in favor of the difference principle gains to those more advantaged are never made at the expense of those less advantaged; instead, any gains to the more advantaged always benefit also the least advantaged, and do so more than any other alternative measure. By contrast restricted utility, even if it provides a social minimum, still permits disadvantages and losses to the worst off so that those better off may prosper; any degree of inequality is allowed in the name of maximizing utility so long as it does not violate the social minimum. Such a situation, Rawls contends, would be morally unacceptable to free and equal persons in a well-ordered society, and is rationally unacceptable to the parties in the original position.

There are different ways to conceive of an economic system based in reciprocity. Even a laissez-faire entitlement system of free transfer and exchange that satisfies Pareto efficiency satisfies reciprocity in a weak sense (assuming the absence of negative externalities) since each transfer makes some people better off and no one worse off. But the Pareto principle and laissez-faire entitlement principles are compatible with enormous gains to the more advantaged while the least advantaged gain only minimally, if at all. This is “trickle-down,” where the poor in effect cannot advance unless the rich maximally benefit, and where benefits to the rich need not benefit the poor at all. The kind of reciprocity provided by the principle of restricted utility is more robust than laissez-faire and the Pareto principle since it guarantees a social minimum. Everyone has a stake in the economic system at least to the degree that it meets the basic economic needs of all. But, beyond this point wealth and income are generated and distributed so as to maximize overall wealth and therewith (presumably) overall utility. Further gains to those better off need not advance the position of the least advantaged, and indeed sometimes may come at their expense so long as the social minimum is maintained.

The “deeper idea of reciprocity” (JF 124) implicit in the difference principle has just the opposite tendencies: the more advantaged may not gain at any point unless their gains benefit the least advantaged and benefit them maximally, better than any alternative arrangement of institutions. The best way to understand the idea of strict reciprocity Rawls incorporates into the difference principle is by referring to figure 6 in Theory (sect. 13) and figure 1 in Justice as Fairness (p.62). The difference principle requires the distribution of powers, prerogatives, and economic resources that put the least advantage on the highest point on the efficient production curve, D, which is the point that is closest to an equal distribution. At D and all prior points on the curve, improvements to the most advantaged are always accompanied by improvements to the least advantaged and vice versa. Hence with all increments to social output, no one gains at any point at the
expense of the other. This relationship of reciprocity does not hold at points to the right of D, where further gains to the more advantaged may increase aggregate wealth and utility, but come at the expense of the less advantaged.

What bearing does this have on choice in the original position? Even if the deeper reciprocity achieved by the difference principle seems morally appealing to us, the parties are not similarly motivated by moral intuitions of fairness. They must be moved to agree on the difference principle for rational considerations alone. So why should the parties in the original position care about the deeper reciprocity achieved by the difference principle? Why wouldn't it be rational for them to agree to a more superficial reciprocity, as allowed by restricted utility, thereby taking a chance that they might be among the affluent in the capitalist welfare state? After all, if they end up among the least advantaged, they may only be moderately worse off than they would have been under the difference principle.

The reasons that speak in favor of the parties' rational choice of the difference principle are their higher-order interest in developing their capacities for justice, their concern for their self-respect, their concern for stability, and the strains of commitment. Compare the difference principle with the principle of restricted utility: Once the social minimum is met, restricted utility does not guarantee that the worse off will benefit in any way from further gains to those better off. Quite the contrary, further gains to more advantaged may even disadvantage the less advantaged—for example, a falling minimal wage rate in the face of an increased supply of labor results in a greater share going to capital, which may benefit owners and middle class consumers but not the less advantaged workers. With restricted utility there is no consistent and continuing tendency toward reciprocity of benefits, for once the social minimum is satisfied the less advantaged are as likely to gain nothing as to benefit from further gains to those better off.

Rawls's conjecture is that in the capitalist welfare state structured by restricted utility, the less advantaged are likely to become dispirited, resentful, and frustrated with their situation, for they know that their well-being is neglected and often intentionally sacrificed so that the majority of citizens may prosper. While stability is maintained among the less advantaged as a modus vivendi, still they are likely to withdraw from active participation in politics and public life; for they justifiably feel left behind by society and no longer see themselves as having a stake in increasing social prosperity or as enjoying a respected position in public life. This all-too-familiar phenomenon in the modern capitalist welfare-state is evident from the striking lack of political participation by the poorest members of our society. It may be that welfare-state capitalism is stable, but it is the stability of indifference or hopelessness among the less advantaged, not stability for the right reasons, which is grounded in equal citizens' affirmation of social institutions out of their sense of justice (PL xlii, 391). Due to their lack of self-respect, and the excessive demands the capitalist welfare-state places on their moral sensibilities and capacities for justice, the least advantaged are unable to willingly affirm the organizing principles of society on grounds of their sense of justice. The principle of restricted utility then places excessive strains of commitment on the worse off, and undermines their sense of self-respect, causing them to be resentful of their situation. Finally, their higher-order interest in the full development and effective exercise of their capacities for a sense of justice are not well served by restricted utility. Because of their interests in fully exercising their moral capacities, their sense of self-respect, and their concern for stability, the parties in the original position cannot in good faith rationally affirm restricted utility and the capitalist welfare state when they have the alternative of choosing the difference principle (cf. JF, 128–129). This seems to be Rawls's main argument for the difference principle from the original position.

Rawls's assumption that the capitalist welfare state is supported by the principle of restricted utility seems to be based on arguments from welfare economics in the last century. In response to Rawls's arguments restricted utilitarians further to the left might reply that the enormous inequalities allowed by capitalism and the capitalist welfare state would not be allowed under a well-ordered society that observes the restricted utility principle. For, other things being equal, the decreasing marginal utility to individuals of money and goods of all kinds favors a more equal distribution of income and wealth. Hence Rawls's assumption that
restricted utility favors a capitalist welfare state that sanctions gross inequalities is mistaken.

The ceterus paribis clause above is all-important here. Even if we grant the controversial assumption that individuals similarly situated gain roughly equal utility from marginal increments of income and wealth, the deterrent effects of equal distributions on (among other things) incentives to work and take risks with productive resources have long been emphasized by economists and more conservative utilitarians in the classical liberal tradition (Hume, Adam Smith, J.S. Mill, Sidgwick, Edgeworth, etc.) These and other questions regarding the kinds of economic systems supported by variants of the principle of utility are empirical questions which cannot be resolved by philosophy.

7. Is the Original Position Necessary or Relevant?

Now that the arguments for the principles of justice have been outlined, this is a good place to consider the objection that the original position is not necessary or even morally relevant. There are at least two notable versions of this argument; the more common one that was initially set forth by Hume will be discussed in the final section. Here I outline a more sympathetic version of the objection by a fellow contractualist, T.M. Scanlon.

As discussed earlier, Rawls depicts two social contracts: rational agreement among interested parties in the original position corresponds to reasonable agreement among members of a well-ordered society motivated by their sense of justice. This correspondence suggests, T.M. Scanlon says, that, rather than the original position, “the point of view which the theory takes as fundamental is actually that of a person in society” (Scanlon, 1975, 177). Scanlon contends that since the arguments in the original position except maximin depend upon agreement among reasonable persons in a well-ordered society, the original position is not necessary. Moreover, the idea of self-interested agreement behind the veil of ignorance distracts from the real justification for the principles of justice—that they are reasonably acceptable and could be justified to persons with a sense of justice in a well-ordered society (Scanlon, 1982, 127). Scanlon's own contractualism relies upon a similar idea of morally motivated persons' agreement on a system of rules for the regulation of behavior that no one could reasonably reject (1982, 110). He abjures the idea of rational choice and agreement by interested parties from an original position or other impartial perspective.

Rawls says that one reason for denying the parties information about themselves and imposing the veil of ignorance is that for a contract theory to rely only on an informal idea of reasonable acceptability and agreement among persons situated in society is less precise and its results less definite. If full information of particular facts were allowed the parties, “only a few rather obvious cases could be decided. A conception of justice based on unanimity in these circumstances would indeed be weak and trivial” (TJ 141). Here one might reply that Rawls is shortchanging the force of his own arguments, from the strains of commitment, publicity, self-respect, and reciprocity, all of which depend on the point of view of reasonable persons in a well-ordered society. Perhaps Rawls thought the maximin argument was needed to complement these arguments. In any case the main objection Scanlon raises seems to be that purely rational choice and agreement based in individuals' interests is not representative of or relevant to moral justification of principles. The (self-) interested choice of rational individuals, even if behind the veil of ignorance, bears little resemblance or relationship to moral judgment and justification among reasonable persons in society. Joshua Cohen has raised similar objections. (J.Cohen, forthcoming)

The answer to this objection (if there is one) is too complex to deal with here. It might begin with Rawls's assumption of the circumstances of justice, his assumption that questions of social justice arise due to the conflict of individuals' legitimate pursuit of their individual interests. Since justice for Rawls concerns the fair terms of social cooperation among persons with different interests and conceptions of their good, he evidently thinks their “essential interests,”--those essential to the rational pursuit of their good, including the moral
powers and primary social goods—should be represented in the agreement on principles of justice that structure their social relations. These interests are especially relevant to agreement on principles of justice for the basic structure, for the primary goods are what these principles distribute.

Second, like Kant, Rawls regards moral principles as the imposition of reasonable constraints on individuals' rational deliberations and choices in pursuit of their interests. This is the “priority of right” over the good (TJ 31f./27f.), which for Rawls represents the structure of practical reasoning of moral persons. As Kant (according to Rawls, LHMP and CP 497ff.) seeks to model the structure of practical reasoning in the categorical imperative procedure, so Rawls seeks to model the priority of right over the good in the original position by the reasonable constraints imposed on rational individuals' choice of principles that promote their good.

Third, Rawls is deeply concerned (again like Kant) with demonstrating that the Right and the Good are “congruent.” Congruence in Theory means, first, that, assuming the most reasonable principles are established, justice and the corresponding sense of justice are compatible with our rational pursuit of our individual interests. Moreover, justice and the exercise of their capacities for justice can be a fundamental good for moral persons, worth pursuing and incorporating into their life plans for their own sake (TJ 398/350 rev.; 567–75/496–505 rev.). (See also, Freeman, 2003, 277–315.) These considerations are all relevant to Rawls's reasons for “modeling” reasonable agreement on principles among members of a well-ordered society by the original position and the parties' rational agreement behind the veil of ignorance.

8. The Original Position and the Law of Peoples

In *The Law of Peoples*, his essay on international justice, Rawls appeals to the original position a second time, to argue for principles of justice that hold among different political societies, or “peoples.” Rawls contends that a theory of social and political justice requires principles of justice to regulate the foreign relations of well-ordered societies with one another. He calls these principles “the Law of Peoples.” Since he conceives of the Law of Peoples as regulating relationships between political societies, he imagines an agreement, not among all the individuals in the world's population, but, first, among the representatives of liberal democratic well-ordered societies. In this agreement among (liberal) peoples, each political society is to be regarded as equal, with equal rights of participation in this agreement; they each have the same number of representatives, no matter the size of their population. To represent the equality of peoples and guarantee fairness of the agreement, Rawls once again utilizes the original position as a hypothetical situation from which representatives of well-ordered liberal peoples decide principles of international justice. The parties to this agreement are once again to be regarded as ignorant of particular facts about their societies, including the size of their population, their natural resources and level of produced wealth, their social and ethnic cultures, and other particular facts, knowledge of which might result in unfair bargaining advantages and lead to an unfair agreement. The representatives of each society are motivated by their fundamental interest in maintaining the justice of their own societies, as this is defined by justice as fairness or some other liberal conception.

Rawls contends that the Law of Peoples that would be agreed to among representatives of “free and democratic peoples” from the original position consists of (at least) eight principles (LP 38). These principles, except for the last requiring a duty of assistance owed to “burdened peoples,” are familiar: They require respect for peoples' freedom and independence; the obligation to observe treaties and undertakings; the equality of peoples; a duty of non-intervention; a right of self-defense and no right to instigate war except in self-defense; a duty to honor human rights; a duty to observe restrictions on the conduct of wars; and finally the duty to assist people living under unfavorable conditions that prevent their having a just or decent political and social regime. Rawls further maintains that since non-liberal but “decent” peoples would also agree to these same principles, liberal peoples have a duty to observe the Law of Peoples in relations with
them, even though decent peoples are not wholly just in their internal organization. But the Law of Peoples, including the duty of non-intervention, do not entirely apply in interactions with “outlaw regimes,” those which do not respect the human rights or good of their own people. Liberal and decent societies may intervene in their internal affairs in order to protect the human rights of their members.

Rawls’s Law of Peoples has been widely criticized since it does not allow for an original position agreement among all the world’s members but only among their national representatives. Nor (a related criticism) does Rawls envision agreement upon a global principle of distributive justice, or even a global tax on more advantaged societies' wealth or natural resources, to be redistributed to less advantaged societies. (See, e.g., Pogge 1989, 2007.) These are complicated issues that cannot be addressed here. (For a defense of Rawls, see Freeman 2007a, ch. 8–9 and 2007b ch.10.) In general, Rawls's positions on these issues are grounded on an assumption of the political and institutional bases of distributive justice, and the fundamental role of society and its basic social institutions in the development of our natural and moral capacities and in determining our characters, aims, and future prospects.

9. Constructivism, Objectivity, Autonomy, and the Original Position

In A Theory of Justice Rawls provides a “Kantian Interpretation” of the original position and the principles of justice (TJ §§ 40, 78). The Kantian interpretation is the first step towards Rawls's Kantian constructivism (CP, ch. 16), and his later political constructivism (PL, ch.3). Rawls says that for Kant, “a person is acting autonomously when the principles of his action are chosen by him as the most adequate possible expression of his nature as a free and equal rational being” (TJ 252/222). What is missing from Kant, Rawls says, is an attempt to show how moral principles express our nature. “This defect is made good [by] the original position” (TJ 255 /224). The original position can be interpreted as a “procedural interpretation” of our nature as free and equal rational beings, and therewith of Kant's conception of autonomy and the categorical imperative (TJ 256 /226). For in the original position, because of the veil of ignorance and other moral constraints, the parties' choice is made “independent of natural contingencies and accidental social circumstances” (TJ 515/451); thus the principles of justice are not chosen “heteronomously,” on the basis of our social position, natural endowments, particular wants, or the particular kind of society we live in (TJ 252/222). Instead, the parties are all represented in the same way, as free and equal rational persons who choose principles of justice subject to all relevant moral conditions. Rawls says that the original position might thus be regarded as incorporating “conditions that best express their nature as free and equal rational beings.” On a Kantian view, our moral nature is defined by our capacities for practical reasoning. These are our capacities to be rational and reasonable. The moral powers are the relevant capacities of practical reasoning in so far as they bear on justice. “Acting autonomously is acting from principles that we would consent to as free and equal rational beings” (TJ 516/453). The description of the original position expresses “what it means to be a free and equal rational being,” and even “resembles the point of view of noumenal selves” (TJ 255–256/225).

In addition to expressing our autonomy, the original position is also objective (TJ 587/514). For it requires that the parties adopt a common impartial standpoint and make a considered rational choice under conditions that require them to abstract from their particular interests and circumstances. Moreover, all are motivated by higher-order interests in their moral powers, which represent their “nature as free and equal rational beings.” Finally, the OP is designed to incorporate all the relevant reasons and restrictions on arguments for principles of social and political justice (TJ 18/16). In so far as the OP is an appropriately defined objective point of view incorporating all relevant moral reasons and conditions on rational choice of principles of justice, and the parties therein come to a unanimous agreement, then the principles agreed to are also objective (TJ 516–517/453). Together with the universality requirement, we can infer from the objectivity of the principles of justice that they apply to and are binding on persons in all societies. As Rawls says of the original position in concluding TJ, “to see our place in society from the perspective of this position is to see it sub specie
Rawls affirms the objectivity and “correctness” of the principles of justice, and he says they are the “most reasonable” principles. But he does not say at any point that the principles of justice are “true.” He does however say that judgments based upon the principles of justice are true, and this applies to both particular judgments and to the moral rules of justice that are required by the principles of justice (CP 355). These claims are implicit in Rawls's Kantian constructivism. (It is noteworthy that Kant himself did not say that the categorical imperative is true; instead he said it is “universally valid.”) In general, constructivism in moral philosophy is a view about the possibility of the correctness of moral principles and judgment. “Realism” says that the correctness of moral judgments resides in their truth, and that truth of most fundamental moral principles consists in their corresponding to a moral order (of moral facts or objects) that is antecedent to reason and to principles of practical reasoning. Objectivity of moral judgment is then defined by realists as judgment made from a perspective of reasoning that is likely to lead to discernment of these antecedent objects of truth. In “Kantian Constructivism in Moral Theory” (Lecture III, CP 340ff.), Rawls discusses a version of realism found in Henry Sidgwick and G.E. Moore, “rational intuitionism,” which he contrasts with moral constructivism.

Kantian constructivism inverts the relationship realists see holding between objectivity and truth; rather than objective judgments being grounded in the discernment of antecedent objects of truth, objectivity (of judgment) precedes the object (of truth). This means that at the level of fundamental moral principles the correctness of these principles depends, not on their correspondence to a prior moral order, but on their following from (or being among) the fundamental principles of practical reasoning. The objectivity of judgment that is involved in reasoning from an objective perspective according to relevant principles of practical reasoning results in objective moral principles that are the bases for judgments of moral truth. The moral facts that are the objects of these moral truths are not then prior to, but are the facts that are singled out as relevant by moral principles and principles of practical reasoning (CP 516).

Among other advantages, moral constructivism relieves moral theory of the burden of having to account for the correctness of moral judgments in terms of their correspondence to a mysterious domain of moral facts (natural or non-natural) that must exist prior to practical reasoning. Constructivism also enables Rawls to provide an account of the objectivity of moral judgment and correctness of moral principles that is consistent with a Kantian idea of autonomy, In setting forth in detail the manner in which the principles of justice are justified on the basis of certain conceptions and principles that originate in practical reasoning itself, Rawls sees himself as having carried through with the Kantian aspiration of showing how moral principles of justice are the result of “reason giving principles to itself, out of its own resources.”

The original position plays a crucial role in Kantian constructivism. Rawls says the original position is a “procedure of construction” that specifies an objective point of view from which to derive principles of justice. Principles of justice are “constructed” on the basis of ideal conceptions of the person and of society (Ideas of reason, in Kant's sense). In Kantian constructivism (CP ch.16), Rawls alters the Kantian interpretation's idea of “our nature as free and equal rational beings,” into an ideal “conception of the person,” regarded as a “free and equal moral person,” with the two moral powers of practical reasoning. It is this conception of the person, along with the ideal of social cooperation represented by a well-ordered society, that is “modeled” or “mirrored” by the original position. In “Kantian Constructivism,” Lecture II, Rawls explains how each of the relevant features of the original position (including the veil of ignorance and the description of the parties as “rational agents of construction”) “represent” in some fashion the ideal conception of the person and of society that underwrite Kantian constructivism. Since the original position presumably represents all the relevant ideas and principles of practical reason, Rawls contends that whatever principles chosen therein are objectively correct, as a matter of “pure procedural justice at the highest level” of practical reasoning. (See below on role of original position in reflective equilibrium.)
Subsequently, in *Political Liberalism*, Rawls “brackets” (he neither affirms nor denies) claims about the nature and possibility of moral truth and universal objectivity of moral judgments. Still, Rawls affirms “for political purposes” the objectivity of the original position and therewith of the principles of justice. Once again, the original position is set up to represent and model an ideal of persons and society. But Rawls detaches the ideal of free and equal moral persons from the Kantian Interpretation, and now regards it as a purely “political ideal” of citizens that is implicit in the political culture of a democratic society. The political ideals of citizens and of society as well-ordered provide the basis for “political constructivism” where the original position again plays its role as a procedure of construction for political principles of justice. Rawls contends that the original position bestows “political objectivity” on principles of justice. Political objectivity is a kind of objectivity that presumably even the moral skeptic can accept, so long as he/she is “reasonable” with an effective sense of justice. It does not imply that the principles of justice apply universally across all social and temporal conditions. Instead, principles of justice are objective in so far as they apply to all reasonable and rational persons who conceive of themselves as free and equal citizens, have a sense of justice, and want to cooperate with others on terms that reasonable persons call all accept.

If we see the original position as a procedure of construction, then perhaps we can gain a clearer idea of the role of the idea a social contract in Rawls's and other arguments for principles of justice. A frequent criticism of social contract doctrines, dating back to David Hume, is that the idea of agreement (hypothetical or actual) does no real work in justifying principles of justice. The justification of principles is said rather to reside in the *reasons* for the parties' agreement; the social contract itself does not add to but only obscures the force of these reasons. Thus Hume argued that Locke's social contract argument is an unnecessary shuffle, and that political obligations and the duty of allegiance cannot be based simply on a promise or agreement. For why should we honor our promises and agreements? It can only be because of their public utility, Hume says. But political obligation is also grounded in the public utility of people respecting political authority and obeying its laws. Hence there is no need to refer to any social contract to ground political obligations, since all have their justification in public utility (Hume, 1748; Rawls, LHPP, 169–173).

To respond: It is certainly true that the reasons motivating the parties in the original position (their freedom to pursue their good, their higher-order interests in developing the moral powers, acquiring an adequate share of primary goods, etc.) as well as the reasons and ideals modeled by the original position itself (the freedom and equality of moral persons, fairness, publicity, the ideal of a well-ordered society, and so on) are fundamental to the justification of the principles of justice agreed to. But this does not mean that the agreement from the original position plays no essential role in the justification of these principles. For the original position itself is designed to be a representation and summary of all the moral reasons and rational interests relevant to justifying principles of justice. Even if it be conceded that rational choice and agreement in the original position, or reasonable agreement in a well-ordered society, do not provide independent reasons for principles of justice, still they serve the crucial, perhaps necessary role of organizing the reasons that the original position incorporates and authorizing inference to the principles of justice.

Consider a parallel argument to the contention that agreement in the original position provides no reasons for and plays no role in justifying principles of justice. We would surely think it peculiar if someone said of a formal proof: “The reasons for the proven theorem are provided by the axioms it presupposes and not by the rules of inference enabling proof.” For rules of inference are indispensable for proof of any deductive conclusion. Agreement in the original position stands in a similar relationship to the principles of justice. Agreement in the original position is not itself a proof from prior premises according to rules of deductive inference. But like rules of deductive inference, agreement in the original position establishes the justifying connection between relevant moral and practical reasons (incorporated into the original position itself) and principles of justice.

Rawls conjectures in *Theory* that the argument in the OP “aims eventually to be strictly deductive….a kind of moral geometry” (TJ 121/104–5; see also JF 17, 83). This is not a concession to Hume's argument, since
Rawls envisions the parties' selection of the principles of justice as part of this conjectured deduction (JF 133). Moreover, even assuming there is a very complicated deductive proof (not what Hume envisioned) of the principles of justice that does not rely on a hypothetical contract. Rawls's attitude would be that the original position argument still serves as a useful heuristic for public reason that enables citizens to understand the enormous complexities of this conjectured proof. Finally, Rawls later qualifies his claim of deductive proof and says that in fact the argument from the original position cannot “fully attain” an ideal of rigorous deductive reasoning; for there are indefinitely many considerations to be appealed to in the original position, and deciding the balance of reasons in favor of the principles of justice depends, not on deduction, but “on judgment informed and guided by reasoning” (JF 133–4).

Finally, what is the relationship between the original position and Rawls's non-foundationalist method of justification, “reflective equilibrium”? Rawls distinguishes three points of view; that of the artificial parties in the original position; of the idealized members of a well-ordered society; and “that of you and me” (PL 28), which is the position from which any conception of justice is to be assessed in reflective equilibrium. Reflective equilibrium begins with our shared “considered convictions” of justice at all levels of generality, from particular to our most general judgments. It then seeks to discover the principles of justice that best “fit” with these considered convictions in a “wide reflective equilibrium,” after considering all the reasonable alternative conceptions of justice. (See TJ § 9; JF § 10.)

Rawls said at one point that “reflective equilibrium works through the original position.” The suggestion is that the original position itself supplies, in large part, the relevant sense of “best fit” of considered moral convictions with principles of justice. The principles of justice that best fit our considered convictions of justice are those that would be chosen by rational persons within this hypothetical-choice “procedure of construction,” which itself “represents” our considered convictions regarding all the relevant reasons for arguments about principles of social justice. Rawls says in concluding A Theory of Justice, “Each aspect of the original position can be given a supporting explanation. Thus what we are doing is to combine into one conception the totality of conditions which we are ready upon due reflection to recognize as reasonable in our conduct towards one another” (TJ 587/514). Thus, in response to the Humean argument that social contract doctrine is superfluous: Until opponents of Rawls's contract doctrine present an an alternative argument for the principles of justice that is more persuasive and inclusive of all the relevant reasons bearing on justice, Rawls's original position argument should be regarded as essential to justification of the principles of justice (in so far as they are justifiable).

Bibliography

Primary Sources

The following works by John Rawls are cited above.


Secondary Sources


——, 1996, Justice and Justification, Cambridge: Cambridge University Press. [see especially Daniels' essays on reflective equilibrium]


**Academic Tools**

- How to cite this entry.
- Preview the PDF version of this entry at the Friends of the SEP Society.
- Look up this entry topic at the Indiana Philosophy Ontology Project (InPhO).
- Enhanced bibliography for this entry at PhilPapers, with links to its database.

**Other Internet Resources**


**Related Entries**

ethics: deontological | justice: international distributive | justification, political: public | liberalism | Rawls, John | reflective equilibrium | social contract: contemporary approaches to

---

Copyright © 2014 by Samuel Freeman <sfreeman@sas.upenn.edu>

Open access to the SEP is made possible by a world-wide funding initiative. Please Read How You Can Help Keep the Encyclopedia Free

Stanford Center for the Study of Language and Information