Introduction: the original position and *The Original Position* – an overview

Timothy Hinton

John Rawls's idea of the original position – arguably the centerpiece of his theory of justice – has proved to have enduring philosophical significance for at least three reasons.[1]

First, it offered a fresh way of thinking about problems of justification and objectivity in political philosophy. At the heart of these difficulties is the need to find an objective point of view from which to deliberate about matters of basic justice. Here "objective" implies "not mired in partiality" and "not biased by one's particular position in the social world." The original position is a hypothetical contractual situation in which parties who are ignorant about crucial features of themselves (such as how wealthy or talented they are, and what their vision of the best way to live is) are to select the principles of justice to regulate the basic institutions of their society. In selecting those principles, the parties are thought of as entering into an agreement that binds them to honor whichever principles they choose. By specifying that the parties are ignorant of matters that would allow them to favor themselves, Rawls vividly and unforgettably captures a widely shared sense that principles of justice cannot be justified by appealing to morally irrelevant considerations.

The original position is important in the second place because of the many interesting philosophical questions it raises. As soon as Rawls's argument had been fully digested, many philosophers felt that something was amiss with it. Questions abound. How could the fact that I would have agreed to certain principles in a special situation of choice give those principles binding authority over me?[2] Is it true that Rawls's two principles in particular are

[1] Rawls's first published paper "Outline" (1951) focused on the question of justification in moral philosophy. He went on to present an early version of his two principles of justice in two papers entitled "Justice as Fairness" (1957, 1958). The idea of the original position as a way of justifying those principles only emerged in his 1967 paper "Distributive Justice."

[2] Rawls himself considers and responds to this objection in *TJ*, at pp. 21 and 587. It was taken up by Ronald Dworkin in his review of *TJ*, "The Original Position," and more recently Habermas raises it.

the most rational choice that could be made in the original position?[3] Are the assumptions needed to get the device off the ground really as weak and untroubling as Rawls seems to have thought?

Finally, the original position is significant because of its evident traction: it has inspired other philosophers to take up alternative positions, to rethink it, and to conceptualize afresh the philosophical problems to which the idea was initially addressed.[4] The vast literature on Rawls's idea – and the use he himself made of it in his subsequent work – are testament to its capacity to inspire further philosophical reflection.

In this introduction, after briefly describing its role in *TJ*, I shall lay out the main features of Rawls's argument from the original position as presented in that text.[5] I will then indicate some of the ways in which Rawls's use of the original position changed in his subsequent work. I shall end with a description of the chapters that make up the volume.

0.1 The place of the original position in *TJ*

In the preface to *TJ*, Rawls tells us that he considers his theory of justice to be a systematic alternative to utilitarianism. The theory Rawls presents – which he names *justice as fairness* – is a form of contractualism he describes as "highly Kantian in nature" (*TJ*, p. viii). What recommends justice as fairness over utilitarianism, according to Rawls, is both that it more faithfully reflects our considered beliefs about justice and that it "constitutes the most appropriate moral basis for a democratic society" (*TJ*, p. viii).

Let's take each of these claims in turn. What exactly are the considered beliefs that Rawls has in mind? They are the claims that first, justice is the most important virtue of social institutions; and second, that justice confers on each person an inviolable status. Together these constitute what he calls "our intuitive conviction" that justice has primacy (*TJ*, p. 4). Rawls wants to know: are we justified in thinking this? And what are the strongest reasons we can give for thinking it? Rawls's theory of justice is meant to answer these questions: it is intended to explain why we are justified in thinking that justice is so basic by uncovering the foundations of that belief.

But we should not lose sight of Rawls's second reason for taking justice as fairness to be superior to utilitarianism. This has to do with the highly

[3] For an examination of some of these issues, see Harsanyi, "Can the Maximin?"
[4] T. M. Scanlon provided a famous alternative form of contractualism in "Contractualism."
[5] For an excellent discussion of the original position, see Freeman, *Rawls*, Chapter 4.

practical nature of Rawls's overall project: he wants to find a moral justification for the institutions that comprise a constitutional democracy, and he wants us – you and me – to find in his theory of justice grounds that could justify our own (presumed) commitment to a democratic society.

If we put these thoughts together, we can formulate a basic question. What conditions must a democratic society satisfy in order to count as perfectly just? The basic hypothesis driving both the argument of *TJ* and much of Rawls's subsequent work is that the traditional idea of a social contract, with appropriate modifications, has a fundamental role to play in answering this question. Rawls wants us to think of the basic principles of democratic justice as ones that would be agreed to by "free and rational persons" who sought to "further their own interests" while they were tasked with selecting those principles in a completely fair choice situation (*TJ*, p. 11).

It is important to bear in mind that the original position is embedded in the broader coherence-driven conception of justification that Rawls named "reflective equilibrium." Rawls is assuming that all our beliefs, including of course our beliefs about justice, form a systematic whole, which, in the ideal case, would be fully governed by rational standards. Justification is a matter of finding the right kind of coherence among the elements that make up this system, including those elements that tell us what it is for our beliefs to be well justified. In the course of seeking an appropriate ordering for our system of beliefs about justice in particular, Rawls says, we need to set out from certain basic convictions that we treat as "provisional fixed points" in our reasoning. As examples of the kind of beliefs he has in mind, Rawls mentions our confidence that "religious intolerance and racial discrimination are unjust" (*TJ*, p. 19). To decide *which* of our considered judgments to take as provisionally fixed, we can set aside those about which we feel hesitant, as well as those made when we are fearful or upset. With these provisional lower-level judgments in place, the aim then is to find more general or abstract principles which, in conjunction with our knowledge of the relevant facts, would support these very same judgments. We could put this by saying that the principles we are after ought to be able to function as the normative major premises in an argument whose conclusions were the basic judgments from which we began.

In the method of reflective equilibrium, the original position functions as an intermediate step. It stands between our initial basic convictions and the more abstract principles to which inference will be made. This is why Rawls insists that it be as fair as possible: in effect, it works to filter out unreliable or biased reasons that might be proposed in support of candidate

principles of justice. It explains why the parties in the original position should be ignorant about such things as their socioeconomic status, their ethical or religious commitments, and their native endowments, including their strength and intelligence.

Because they are ignorant about these matters, Rawls describes the parties in the original position as being behind a "veil of ignorance" (*TJ*, p. 136). In addition, they are understood as enjoying perfect equality: none has any more information or bargaining power than any other. Furthermore, the parties are neither altruists nor egoists, nor are they spiteful or malicious. As Rawls puts it, they have no "interest in one another's interests" (*TJ*, p. 148). Crucially, they are rational: they want to do as well for themselves as they can, given their situation. As a result, they want to choose whichever set of principles will best advance their interests. This presupposes that they know they have certain interests of their own: they want to ensure that, once the veil of ignorance is lifted, they will be in a position to live out their lives as best they can. They are trying to find principles that will enable them to meet these goals. But since they do not know what their own values and life plans are, they are forced to make a choice under uncertainty.

Of course, Rawls wants us (you and me, here and now) to be persuaded that the set-up of the original position makes sense. This is why he suggests that we will agree with him on the most reasonable way to go about creating the original position. In other words, Rawls hopes that we will agree that the constraints to which the parties in the original position are subject are ones that we ourselves find reasonable. Rawls hopes that even if we do not initially accept these conditions,

> perhaps we can be persuaded to do so by philosophical reflection. Each aspect of the contractual situation can be given supporting grounds. Thus what we shall do is to collect together into one conception a number of conditions on principles that we are ready upon due consideration to recognize as reasonable. (*TJ*, p. 21)

But it is important to bear in mind that Rawls insists that the stipulations that set up the original position are all open to reflective re-examination, and, where necessary, to rational revision:

> We begin by describing it so that it represents generally shared and preferably weak conditions. We then see if these conditions are strong enough to yield a significant set of principles. If not, we look for further premises equally reasonable. But if so, and these principles match our

> considered convictions of justice, then so far well and good. But presumably there will be discrepancies. In this case we have a choice. We can either modify the account of the initial situation or we can revise our existing judgments, for even the judgments we take provisionally as fixed points are liable to revision. By going back and forth, sometimes altering the conditions of the contractual circumstances, at others withdrawing our judgments and conforming them to principle, I assume that eventually we shall find a description of the initial situation that both expresses reasonable conditions and yields principles which match our considered judgments duly pruned and adjusted. (*TJ*, p. 20)

The parties in the original position are trying to choose the most basic principles that will regulate their society, that is, the basic principles of justice. Rawls argues that in this choice situation, the parties would choose the following pair of principles:

> The Equal Basic Liberty Principle: Each person is to have an equal right to the most extensive system of equal basic liberties compatible with a similar system of liberty for others. (*TJ*, p. 250)

> The Second Principle (which comes in two distinct parts, and in which the first part has lexical priority over the second): (a) The Fair Equality of Opportunity principle: Social and economic inequalities are to be arranged so that they are attached to offices and positions open to all citizens under conditions of fair equality of opportunity; (b) The Difference Principle: Social and economic inequalities are to be arranged so that they work to the greatest expected benefit of the least advantaged group in society. (*TJ*, p. 83)

0.2 The core of the original position argument

In my discussion of the argument by which Rawls derives these principles from the original position, I shall, for the sake of simplicity, consider the case in which the parties are making a single pairwise comparison.[6] Here, they are choosing between Rawls's two principles and

[6] In both *TJ* and *JFR*, Rawls presents the argument in two distinct stages. In the latter text, he describes these as "the first fundamental comparison," in which his two principles are paired with the principle of average utility, and "the second fundamental comparison," in which his two principles face off against "an alternative exactly the same as those principles except in one respect. The principle of average utility, combined with a suitable social minimum is substituted for the difference principle" (*JFR*, p. 120).

> The principle of average utility: All social goods – such as rights, opportunities, income and wealth – are to be distributed in such a way as to maximize the average utility of the members of society.

In his first main line of argument, Rawls suggests that it is useful to think of his two principles as "the maximin solution to the problem of social justice" (*TJ*, p. 152). This involves an appeal to

> The maximin rule: Choose that alternative whose worst outcome is better for you than the worst outcome of any other alternative (in a slogan: Maximize the minimum).

Rawls's intention is to characterize the original position in such a way that it is rational for the parties to employ this rule in making their choice. Once this has been done, he says, somewhat optimistically, "a *conclusive* argument can . . . be constructed for these principles" (*TJ*, p. 153, my emphasis).

Rawls believes that it is rational to use the maximin rule in making a decision under uncertainty if the following conditions are met:

1 there is strong reason for discounting the probabilities attached to the various possible outcomes;
2 one has a conception of the good involving a threshold ***t*** which is such that it is much better to be at ***t*** than to be anywhere below ***t***; however, once ***t*** is achieved, further gains above ***t*** either have very little or no significance at all;
3 the situation involves "grave risks": the outcomes one rejects by using the rule are very bad indeed.

In making a case that condition (1) holds in the original position, Rawls reasons as follows. Since the parties do not know any specific facts about their society, they have no way to determine how likely it is that they will end up in any given position. (For example, they have no idea what the possible social positions are, nor do they know how many people occupy any particular position.) Rawls also emphasizes here that the parties (as individuals with continuing lines of descendants) have to be able to justify their choice to others.

In arguing that condition (2) is satisfied in the original position, Rawls appeals to the fact that his two principles seem to guarantee a "satisfactory minimum," while the principle of average utility does not. The first principle guarantees basic liberties for all, while the second principle gives the least advantaged as much as it is possible for them to have. Of course, some people (say, those at the top end of the income scale) might do better in a utilitarian

society than in a Rawlsian one. But, being ignorant about where one will end up makes it rational, Rawls insists, to secure a decent minimum in liberties, opportunities, and income and wealth, rather than taking a gamble in which one might end up much worse off.

Finally, when it comes to condition (3), Rawls argues that none of the social positions that would be permitted by his two principles would turn out to be intolerable, while the possibilities that they rule out would clearly be so. After all, among the possibilities allowed by the principle of average utility could be slavery and serfdom.

In his second main original position argument in *TJ*, Rawls considers what he calls "the strains of commitment." Recall that the parties know that they have a sense of justice. Knowing that they will see themselves as bound by whichever principles they end up choosing, they know that they "cannot enter into agreements that may have consequences they cannot accept" (*TJ*, p. 176). So each of the parties must consider the impact which any chosen principles will have on the lives of people in the society to which they apply. Each must be able to say sincerely: I could commit myself to acting on these principles irrespective of where I ended up as a result of their workings.

With this in mind, Rawls suggests that it would be easier to commit to his principles than to the principle of average utility because, in doing so, the parties would thereby rule out the possibility of having to live with "the worst eventualities" (*TJ*, p. 176). Indeed, he wonders whether anyone could rationally commit themselves to abide by principles whose outcome might involve a loss of important freedoms simply in order to secure greater utility for other people.

Another significant feature of this part of Rawls's argument is its invocation of the importance of publicity in choosing a conception of justice. In the context of the original position, Rawls connects this with the suggestion that what he calls a "stable" conception of justice is a more desirable choice. He explains that such a conception is stable when public awareness of its having been fully realized in a society will lead its members to develop a desire to act in accordance with its principles. Rawls suggests that the parties will be moved by facts about ordinary human motivation to see that utilitarianism "seems to require a greater identification with the interests of others than the two principles" (*TJ*, p. 177). This level of identification, he thinks, would place undue strain on one's ability to remain faithful to the principle of average utility.

There are two further features of the publicity condition. The first appeals to self-respect. Rawls believes that having a lively sense of your own worth is a condition of achieving your ends in life. Unless you are confident that other

people respect your standing in society, you are unlikely to feel that your projects are fully worthwhile. A democratic polity shows proper respect for its members, Rawls says, when it ensures that they have sufficient access to what he calls the social bases of self-respect.

When a society follows the two principles of justice, each person's advantage is taken into account and each takes part in a cooperative scheme of mutual benefit. Since each knows that his or her good is being protected, each can have confidence that he or she is being treated as an equal. This connects with the idea of an underlying moral equality between people that is part of the deep structure of Rawls's view.

Rawls's appeal to publicity in his original position argument also invokes the Kantian principle that people are ends in themselves and ought never to be treated as mere means. Rawls gives this a contractualist gloss. On his view, we treat someone as a mere means just in case we treat her in ways that she would not have consented to being treated had she been asked for her consent in a perfectly fair choice situation. Although it is not entirely clear what this comes to, perhaps we can take Rawls's line of thought to be something like the following. To treat others as ends in themselves is to be willing to forgo a larger share of goods like rights, opportunities, and income and wealth, if having that larger share oneself would mean that someone else would end up with fewer of those goods than anyone needs to have.

Recall that, for Rawls, what matters most when it comes to justifying principles of justice is our status as free and equal persons, each with a conception of the good and a capacity to act on the principles of justice. When I view the social world from the perspective of the original position, Rawls holds, I see only free and equal persons engaged in social cooperation. I would want that social world to be one in which no one could be better off than anyone else simply as a result of factors which are arbitrary from the point of view of justice. The factors which are arbitrary in this way include people's social class, their natural talents, and the values and plans that animate their lives. This conception of the original position – which Rawls calls its "intuitive" understanding (*TJ*, p. 22) – is tied to the very moving lines with which Rawls ends *TJ*:

> to see our place in society from the perspective of this position is to see it *sub specie aeternitatis*: it is to regard the human situation not only from all social but also from all temporal points of view. The perspective of eternity is not a perspective from a certain place beyond the world, nor from the point of view of a transcendent being; rather it is a certain form

of thought and feeling that rational persons can adopt within the world. And having done so, they can, whatever their generation, bring together into one scheme all individual perspectives and arrive together at regulative principles that can be affirmed by everyone as he lives by them, each from his own standpoint. Purity of heart, if one could attain it, would be to see clearly and to act from grace and self-command from this point of view. (*TJ*, p. 578)

0.3 The original position after *TJ*

To understand the main shifts in Rawls's thinking about the original position in his writings after *TJ,* it is helpful to bear in mind an important feature of the broader reflective framework in which the device is embedded. Recall that in *TJ* Rawls wanted his conception of justice to clarify and, if necessary, to correct our basic convictions about justice. By doing so, he hoped to justify those convictions, and to help make sense of our commitment to a democratic society.

In that work, however, Rawls thought of the process of achieving these goals as broadly speaking Socratic. For there, the original position was deployed in a fundamentally first-personal search for principles that one ought to accept. This goal is Socratic mainly because Rawls took the search for principles of justice to involve critical self-reflection on one's own beliefs – the kind of reflection in which neither one's initial judgments nor the subsequent principles to which inference is made are to be treated as sacrosanct. One should stand ready to revise either or both under the pressure of sound reasons. So, framed by the Socratic question, the point of the original position is to enable us to figure out what each of us taken individually should believe about justice.[7]

Beginning with the Dewey and Tanner Lectures, delivered in the early 1980s, Rawls set out in a new direction, in which the aim of justice as fairness was somewhat refigured, and, as a result, a second, non-Socratic method come to the fore.[8] What begins to dominate Rawls's thinking is a focus on certain deep and important political difficulties confronting contemporary democracies, difficulties which he hopes his theory can address, and quite

[7] Compare Scanlon's remarks on "Rawls" in Freeman, *Cambridge Companion*, p. 142.

[8] The Dewey Lectures, delivered in April 1980 and published in the *Journal of Philosophy* 77 (1980), 515–72 as "Kantian Constructivism," were re-written for *PL*, appearing there as Lectures I–III. Rawls presented the Tanner Lectures in April 1981; they appear in *PL* as Lecture VIII.

possibly, resolve. This political focus took two main forms during the later part of Rawls's career.

In its first iteration, the practical worry that exercised Rawls concerned the competing claims of freedom and equality in a democratic society. As a result, he adapted justice as fairness to work out a view of justice

> which is congenial to the most deep-seated convictions and traditions of a modern democratic state. The point of doing this is to see whether we can resolve the impasse in our recent political history; namely, that there is no agreement on the way basic social institutions should be arranged if they are to conform to the freedom and equality of citizens as persons. (*PL*, p. 300)

As this passage suggests, Rawls began to employ the original position as part of what might be called an interpretivist method, where the device is put to work in a program of identifying principles of justice that count as the best overall interpretation of the beliefs, practices, and traditions of modern liberal democracies.

In its second iteration, the basic practical worry that shaped Rawls's revised application of the original position was what he called the fact of reasonable pluralism.[9] This is the idea that a plurality of different conceptions of the good is an enduring feature of any democratic society. This is related to an explanatory claim that Rawls advances, namely that such a plurality exists because of the nature of human reason: when human beings think about the ultimate questions of life (like whether or not there is a god or what the meaning of human life is) they are bound to reach different answers. The questions are hard to answer and human reason is such an imperfect thing that we cannot but arrive at different positions. Provided that there is no state-sanctioned religion which all citizens are forced to profess, this kind of plurality is inevitable. A free society always produces it. But, Rawls wants to stress, many of the doctrines that people formulate are "perfectly reasonable": not only can reason not settle which is of them is correct, but in fact many of these views involve no errors from the standpoint of reason (they involve no errors that a reasonable person would never make).

Along with the shift to interpretivism, Rawls began to describe the original position as a *device of representation*, by which we can take him to mean two things. First, the parties to the agreement are now to be thought of as

[9] For some of Rawls's key thoughts about the fact of reasonable pluralism, see *PL* introduction and Lecture I.