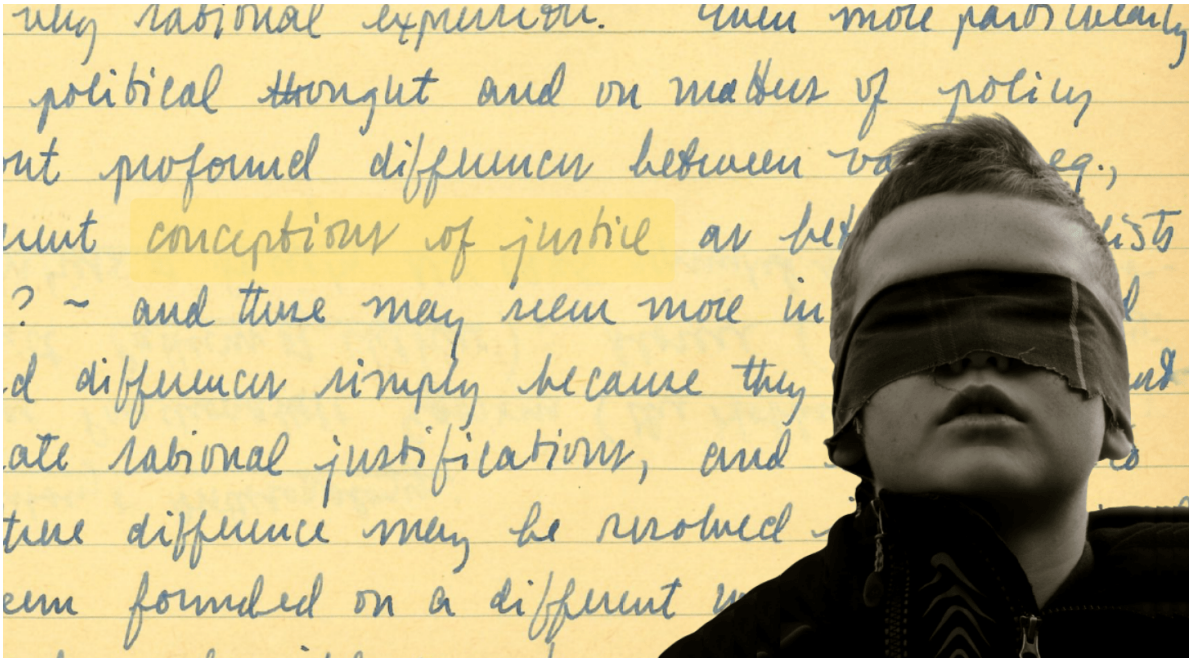
THE ORIGINAL POSITION AND PRINCIPLES OF JUSTICE

Equality & Justification: The Original Position Reconsidered

Véronique Munoz-Dardé



Manuscript from the John Rawls Papers (Harvard University Archives). Graphic design: Maria Oliva Campabadal (CETC).

The original position has an elegance and power beyond most philosophical pictures. It has captured the attention of readers across the world through many generations of students, and is famous well beyond philosophical circles. Yet, as renowned as the original position has become, it is also typically misrepresented and misunderstood. In particular, John Rawls' method of reasoning behind the veil of ignorance is frequently presented as drawing a conclusion mandated by rational choice theory. My aim, in this brief note, is to clarify the main purpose of the original position and to articulate its main defining features in contrast to this dominant misreading.

What Is the Original Position?

The original position is a thought experiment which is designed to express in the simplest, and also the most compelling, way the considerations which bear on our deliberation about principles of social justice. Rawls argues that, taken together, these considerations lead to his two principles of justice.

The original position is intended to constitute a fair and impartial perspective on the question of justice. In it, we are to imagine parties conceived as free and equal persons in ignorance of their individual social, natural and historical characteristics. They do not know their particular conception of the good life, their place in society, class position, social status, race or ethnic group, sex, intelligence, natural endowments, and so on —they are behind a *veil of ignorance*—.

The parties are behind the veil of ignorance, but still aware of social and natural facts which affect human societies. They envisage different sets of principles under which they, and their descendants, will have to live their whole life. Those principles are to apply to the main political and social institutions which together distribute the core benefits and burdens of social life (the *basic structure* of society). In particular, parties consider two alternatives: a utilitarian conception of justice according to which «a society is properly arranged when its institutions maximize the net balance of satisfaction» [1], and Rawls' two principles.

Rawls' two principles of justice stipulate that (a) each person has the same indefeasible claim to a fully adequate scheme of equal basic liberties, which scheme is compatible with the same scheme of liberties for all; and (b) social and economic inequalities are to satisfy two conditions: first, they are to be attached to offices and positions open to all under conditions of fair equality of opportunity; and second, they are to be to the greatest benefit of the least-advantaged members of society (the difference principle) [2].

Rawls contends that reflecting under the constraints of the original position, and thus having to consider seriously what would happen if they were among the least favoured members of society, parties would protect themselves against being badly off. He further argues that his two principles of justice would be adopted over utilitarianism. Parties, that is, would reject principles under which the liberties, opportunities or welfare of some individuals could be sacrificed for the sake of a greater good jointly enjoyed by others. The least well-off would be better off under Rawls' two principles than under utilitarian principles: and that is our reason for choosing them.

> The original position is a thought experiment which is designed to express the considerations which bear on our deliberation about principles of social justice

Since the publication of *A Theory of Justice*, there have been countless arguments to the effect that parties behind the veil of ignorance would endorse principles other than those that Rawls formulates. Such criticisms seem to miss the role that Rawls explicitly assigns to the original position. They treat it as a kind of experiment in which we discover an unobvious conclusion from the principles by which we frame the position. So framed, the original position is supposed to *provide evidence* in favour of Rawls' principles of justice. However, Rawls insists that he has framed the original position in order to *illustrate* the

principles and their rational appeal. Rawls might be mistaken about what the original position leads to, or he might better have chosen a different position from which to start. But following Rawls' intentions, we should expect a harmony between the original position and the principles Rawls recommends.

Why Deliberate from the Original Position? Disagreements About Justice

What is the need for the original position? The original position is conceived to test whether the two principles can be the object of agreement by all members in societies characterised by deep disagreements about what is just and unjust.

«The intuitive idea of justice as fairness, Rawls writes, is to think of the first principles of justice as *themselves* the object of an original agreement in a suitably defined initial situation» [3]. If we isolate the kind of considerations that can justify claims about justice, we can arrive at a shared, public, standard to resolve claims against basic social institutions. Like many of his contemporaries, Rawls is deeply concerned with substantive political and moral disagreements, the presence of which had been emphasized in his own times by the controversies over civil rights, affirmative action, abortion, sexual freedom and the Vietnam War. On the other hand, he also shares the relative optimism of many citizens who think that it is conceivable for a society to be 'well-ordered', in the sense that the shared «public sense of justice makes their secure association possible» [4].

In setting up deliberation about principles of justice, Rawls suggests that it is fruitful to start from the shared yearning for a just society. He holds that despite ongoing disputes about political principles, citizens all understand the need for justice. Citizens have different *conceptions* of justice and disagree about what is just and unjust, and about sets of principles of justice. However, these conceptions have in common an appreciation of the need for justice. There is, therefore, an underlying agreement on the *concept* of justice, defined by this need:

> «[I]t seems natural to think of the concept of justice as distinct from the various conceptions of justice and as being specified by the role which these different sets of principles, these different conceptions, have in common. Those who hold different conceptions of justice can […] still agree that institutions are just when no arbitrary distinctions are made between persons in the assigning of basic rights and duties and when the rules determine a proper balance between competing claims to the advantages of social life».
>
> Rawls, J. (1999) A Theory of Justice. Revised edition. Cambridge Mass.: Harvard University Press, 5 [emphasis added].

Searching for an agreement on principles of justice in a society characterised by deep moral disagreements, therefore, we start from what reasonable participants to a conversation on social justice can, at least on reflection, agree on. The object of this minimal initial agreement focuses in particular on the following reasonable claims: (a) that there should be *no arbitrary distinctions* between people when it comes to fundamental rights and duties, and (b) that the advantages of social life, what Rawls calls 'the fruits of social cooperation', ought to be distributed in a *fair way*.

Thus, the original position isolates some shared, reasonable considerations that citizens *already endorse*: near-platitudes about what constitutes a fair and reasonable society, such as that it would be unfair to discriminate against people for reason of their gender, race or ethnicity.

> The original position isolates some shared, reasonable considerations that citizens already endorse about what constitutes a fair society, such as that it would be unfair to discriminate people for reason of their gender, race or ethnicity

Rawls' purpose is to highlight how, combined together, these moral convictions lead to principles that grant liberties, equality before the law, decent conditions and prospects in life, as well as bases of self-respect to all citizens. He is well aware that many people will find the conception of justice summarized in his two principles highly controversial. However, he wants to argue that claims that constitute widely shared premises between his interlocutors *lead* to these two principles. If you hold the fairly uncontentious claims above, then you are committed to their necessary effects. Hence the introduction of the veil of ignorance and deliberation in the original position. The result, Rawls hopes, is to clarify not only the force of considerations for his own conception of justice, but also the nature of moral disagreements with other positions.

The Controversial Role of Rational Choice

Anybody who has read *A Theory of Justice* may suspect that I have so far radically underplayed the role of rational choice in the original position, and they would of course be right about that.

Moreover, there are in the secondary literature entirely rational choice-based descriptions of the original position. From that perspective, the *point* of the original position is as a set-up for a decision-theoretic choice. A pure and recent example of this perspective is the characterisation of Rawls' original position by Lara Buchak [5]. Buchak conceives of the original position as framing a decision-theoretic choice regarding risk-weighted expected utility of different social policies. She describes the original position thus:

> «[I]ndividuals consider their preferences about institutional arrangements in the “original position,” in which decisions are made behind a “veil of ignorance,” where people do not know ahead of time their “place in society, their class position or social status, their place in the distribution of natural assets and abilities, their deeper aims and interests, or their particular psychological makeup”. Thus, individuals consider their preferences about gambles which correspond to social distributions and in which the possible ‘states of the world’ specify which place each of them will occupy in society».
>
> Buchak, Lara, ‘Taking risks behind the Veil of Ignorance, Ethics, April 2017, 625 [emphasis added].

What the difference between these exegetical strategies brings out is that there is a considerable tension between different conceptions of the original position in the reception of Rawls’ theory. It is fair to say that the controversy between these incompatible conceptions is a consequence of Rawls’ own equivocations regarding the role of rational choice in the deliberation on principles in *Theory*.

Rawls’ equivocation, at least in the initial version of *Theory*, is between a purely rational choice account of the theory of politics in which we prove to self-interested individuals that it is in their prudential interest to adopt certain principles, and a contrasting account of political theorizing within a broader ethical perspective which presents the original position as merely keeping track of reasons that we already have, starting from a shared interest in the need for just institutions. The latter treats it as a thought experiment in which we rule out considerations which would introduce distortion or bias in thinking about social justice, rather than placing any emphasis on self-interest or an axiomatizable conception of the rational. However, the obvious question if one opts for this second interpretative option concerns the role, if any, of rational choice.

What then is the place of rational choice in deliberation about principles of justice? The first thing to note is that Rawls himself changed his mind after the publication of *Theory*: from considering that the theory of justice was *part* of the theory of rational choice to thinking of rational choice as a mere set of formal tools within a conception of justice focused on reasonable principles, rather than on rational interest. Thus, he remarks in the posthumously published *Restatement*:

> «Here I correct a remark in Theory…, where it is said that the theory of justice is a part of the theory of rational choice. … [T]his is simply a mistake […] What should have been said is that the

> account of the parties, and of their reasoning, uses the theory of rational choice (decision), but that this theory is itself part of a political conception of justice, one that tries to give an account of reasonable principles of justice. There is no thought of deriving those principles from the conception of rationality as the sole normative concept».
>
> Rawls, J. (2001) Justice as Fairness: A Restatement, edited by Kelly, E. Cambridge, Mass: Harvard University Press, 82 [emphasis added].

What the mature Rawls highlights is that there is a question about the perspective of the self-interested person we imagine deliberating behind the veil of ignorance, and how they come to accept principles of justice. The question, to put it in T.M. Scanlon's terms, is whether the self-interested person behind the veil of ignorance is held to accept a principle of justice «because he judges that it is one *he could not reasonably reject whatever position he turns out to occupy*, or whether, on the contrary, it is supposed to be acceptable to a person in any social position *because it would be the rational choice for a single self-interested person* behind the veil of ignorance» [6].

Some of the concerns about justice and reasonableness when read simply as facts for rational choice in deliberation should leave an agent cold: they do not compel one to cooperate or not. The original position can be seen as a suggestion of how we can add certain further constraints to rational choice theory such that it becomes rational for the agent to be playing a game with a certain political content.

At various points, Rawls seems to have the ambition of demonstrating how one can derive substantive ethical principles, those of the political domain, from such broadly rational considerations. By ruling out unfairness and arbitrariness, we can find principles of political life that each of the citizens can justify to others. «The aim —Rawls writes— is to rule out those principles that it would be rational to propose […] only if one knew certain things that are irrelevant from the standpoint of justice» [7]. But by the time of his mature reconsideration of his theory, this ambition seems as we just saw to have been given up.

> The question is whether the self-interested person behind the veil of ignorance is held to accept a principle of justice because it is reasonable or because it is rational

One could argue that the costs of the move away from rational choice, and the focus on

reasonableness in the mature work, is precisely accepting that one cannot find suitably general formal constraints from which one can derive the substance of the political principles. But the distinctive advantage of abandoning this ambition is to remove any ambiguity in the fundamental formulation of the scope of the theory, and to focus instead on justifying to others the terms of social cooperation that it would be unreasonable to reject.

Combining into One Conception Conditions of Reasonable Conduct

What is ruled in to the original position are «widely accepted but weak premises» [8]. This mode of deliberation and justification of principles of justice is summarized in the last pages of *Theory*:

> «[J]ustification is argument addressed to those who disagree with us, or to ourselves when we are of two minds. Being designed to reconcile by reason, justification proceeds from what all parties to the discussion hold in common».
>
> Rawls, J. (1999) A Theory of Justice. Revised edition. Cambridge Mass.: Harvard University Press, 508.

To return to a point made above: reasonable citizens may, for example, all agree that racial discrimination is unjust, but be less certain of what distribution of wealth and authority is just. This may lead to fierce disagreements. Faced with uncertainty and disagreement, the original position embodies conditions of non-arbitrariness and of rejection of unfairness that we can on reflection all endorse as reasonable. It allows us to envisage different conceptions of justice and respective sets of principles, starting from shared premises. Using it we can make up our own mind about justice, and do so through deliberating with others.

Through reasoning from the original position, we formulate principles which we can all find justified. The parties are neither people in a just society, nor people in *our* society. They are merely the artificial creatures created for the purposes of our deliberation when we *disagree* about what justice entails, yet are motivated to find principles of justice we can justify to each other.

The point of underlining the ambiguity introduced by considerations of rational choice is that we must be careful not to overestimate the role of the original position in Rawls' overall argument. Rawls conceives of the original position not as a method of discovery of new principles. The purpose of the original position is to justify principles which are *already* available to us, and which in the book are formulated before the argument from the original

position is expounded. In constructing the original position, Rawls is really just highlighting to us the relevant structure of reasons we already care about when we are concerned with justice.

Is this reading too deflationary? I would suggest not: the appeal is great of conceiving of the original position as no more than a brilliant heuristic to deliberate on principles of a just society *starting from* the value people do set on the fact that their institutions are justifiable to others. However, this does raise a further question: Why not present these arguments without bothering with such a thought experiment introducing undesirable ambiguities? This is precisely the lesson that T. M. Scanlon derives from the observation quoted above.

But we are not compelled to follow Scanlon here, and abandon the heuristic. Recall Rawls' point that, in thinking about the justice of our political institutions we face a difficulty, namely, to find a point of view «removed and not distorted by the particular features and circumstance of the existing basic structure» [9]. The principles of justice are principles for the basic structure. However, our own basic structure is unjust. Even when we are motivated to find principles and institutions that are justifiable to each other in the political domain, there are some injustices that we find natural. In other words: even under the deflationary reading above, the original position is philosophically revealing. As Rawls writes at the very end of *Theory*, the original position allows us to envisage our social world and political institutions from the right perspective:

> «[W]e may remind ourselves that the hypothetical nature of the original position invites the question: why should we take any interest in it, moral or otherwise? Recall the answer: the conditions embodied in the description of this situation are ones that we do in fact accept. Or if we do not, then we can be persuaded to do so by philosophical considerations of the sort occasionally introduced.
>
> 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 that we are ready upon due reflection to recognize as reasonable in our conduct with regard to one another (§4).
>
> Once we grasp this conception, we can at any time look at the social world from the required point of view».
>
> Rawls, J. (1999) A Theory of Justice. Revised edition. Cambridge Mass.: Harvard University Press, 514.

Faced with the choice between the different construals of the original position, it might be tempting to suppose that the one which reads rational choice into the essence of the deliberation is to be preferred, because it offers us a more ambitious reading of the *Theory,* articulating the goal of grounding part of collective ethics in rational self-interest. I've tried to sketch here the reasons for supposing that the other reading is in fact the more interesting and ambitious one. Leaving aside the ancient hope of grounding the moral in the prudential, it invites us to see how substantive the principles we must agree on are, just on the basis of finding common concern in reasonable disagreement.

REFERENCES

1 — Rawls, J. (1999) *A Theory of Justice*. Revised edition. Cambridge Mass.: Harvard University Press, 20.
2 — Rawls, J. (2001) *Justice as Fairness: A Restatement*, edited by Kelly, E. Cambridge, Mass: Harvard University Press, 42-43.
3 — Rawls, J. (1999) *A Theory of Justice*. Revised edition. Cambridge Mass.: Harvard University Press, 102 [emphasis added].
4 — Ibídem, 5.
5 — Buchak, Lara, 'Taking risks behind the Veil of Ignorance, *Ethics*, April 2017.
6 — Scanlon, T. M. (2003) 'Utilitarianism and Contractualism', in *The Difficulty of Tolerance.* Cambridge: Cambridge University Press, 146 [emphasis added].
7 — Rawls, J. (1999) *A Theory of Justice*. Revised edition. Cambridge Mass.: Harvard University Press, 16-17.
8 — Ibídem, 18.
9 — Ibídem, 15.

Véronique Munoz-Dardé
Véronique Munoz-Dardé is Professor of Philosophy in the University College London Department of Philosophy and Mills Adjunct Professor of Philosophy at University of California, Berkeley. She is known for her works on ethics and political philosophy. In recent years, she has written articles on the importance of numbers in practical reasoning, on the political ideal of equality, on responsibility and on distributive justice. She is the author of the books *La justice sociale* (2001) and *Rawls: justice et critique* (2014).