The original position: A 'stark fiction' in Rawls's theory of justice

Brown University, December 2011
Jonathan Sozek

No idea is more closely associated with the work of John Rawls than that of the original position (the OP). But what kind of idea is it, and what work does it do? Rawls himself characterizes it variously in his major works as a 'thought-experiment', a 'device' in several senses and even once, in his final book, as a 'play'. This last is not out of place, for as Rawls develops it the OP does exhibit a dramatic structure: in the parties we have *characters*; in their task of selecting a conception of justice under severe conditions, a *tension*; in their deliberations, a *plot*; in their decision and the well-ordered society it yields, a *resolution*. Even, to those of us willing to suspend our disbelief in this 'hypothetical and nonhistorical'[1] situation, Rawls promises a kind of *catharsis* – a new 'seeing' of our place in society from 'the perspective of eternity', carefully defined[2].

I argue in these pages that the OP is a 'stark fiction' at the heart of Rawls's theory of justice – or perhaps better, *under its hood*, for without this dramatic element, 'justice as fairness' (the name Rawls gives to his theory) would be as static and monological an exercise as the utilitarianism it aims to supplant; it would be unable to form citizens. The notion of stark fiction I borrow from Bernard Williams, who identifies as its paradigm example the tragic drama of Sophocles. I begin by examining Williams' account in order to propose that, on Williams' own terms, this notion may be extended in two ways to make possible its fruitful application to

[1] PL, p. 24.
[2] TJ, p. 514. Rawls writes: 'The perspective of eternity is not a perspective from a certain place beyond the world, nor 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.'

Rawls's idea of the OP (§1). Next, with this extended notion of stark fiction in mind, I survey Rawls's presentations of the OP in his major works: first by reconstructing his presentation in *A Theory of Justice*, then by considering the significance of his later, new description of it as a 'device of representation' (§2). In conclusion, I make a case for my reading of the OP as a stark fiction (so, 'what kind of idea is it?') and consider what role it can be seen to play from this new perspective in Rawls's theory of justice (so, 'what work does it do?').

My claim shall be that the OP's dramatic character enables it to function as a 'mediating idea'[3] in justice as fairness, and this in a sense somewhat more broad than Rawls seems to have intended. As a stark fiction, I argue, one may regard the OP as the single feature of Rawls's theory to make possible on a large scale – to *mediate*, for every member of society – a transition from an everyday point of view, so marked by bias and heteronomy, to the point of view of a *citizen*, able autonomously to adopt as her own the moral principles of her society (if they are rational and reasonable). Reading the OP as a stark fiction exposes this transformative and very democratic moment in justice as fairness more clearly than other readings do, and so, I argue, could help spark new conversations between Rawlsian political liberals and those who remain skeptical of his project.

§1 – Williams on stark fiction

Williams introduces his notion of stark fiction near the end of his essay '*The Women of Trachis*: Fictions, Pessimism, Ethics'. The piece is characteristic of Williams' work, and would seem at first to offer little help in making a new reading of Rawls. '[M]oral philosophy,' we are told, 'is still deeply attached to giving good news'[4]. Even if we no longer take seriously

[3] PL, p. 26.
[4] Williams (2007), p. 49.

Leibniz's 'cosmic cost-benefit analysis'[5] or Hegel's story about 'the complex working of the Geist'[6], our moral philosophy remains guilty of trying to 'withdraw our ethical interest from both chance and necessity' as so much 'bad news'[7]. The important question, Williams thinks, is 'how, if at all, moral philosophy *within its own limits* can conduct itself less evasively'[8]. The italicized clause is important: the aim is to extend moral philosophy's 'defective consciousness' in a way we today can take seriously, not by positing some new myth or background of, as Williams puts it elsewhere, 'magical belief'[9].

So Williams turns to fiction. It is a 'traditional idea', he says, 'that fiction can yield salutary *exempla* of virtue and vice'[10]. If one regards the project of defining virtue and vice as a more or less straightforward affair, as Williams thinks moral philosophy has been wont to do, then 'the role of fiction will be that of an efficient aid' – of providing so many 'illustrations'. If, however, one acknowledges a 'certain degree of ambiguity' in the moral life, things look different: '[F]ictions will come to do things that direct statement cannot do, and working through the fiction will itself represent an extension of ethical thought, and conceivably of ethical experience'[11].

---

[5] Ibid.
[6] Ibid., p. 51.
[7] Ibid., p. 54.
[8] Ibid. Emphasis mine.
[9] For more on this ambition see Williams' *Shame and Necessity* (Berkeley: U of California P, 1993). An important motive for Williams in that work is to demonstrate, as he puts it on p. 62, that although many arguments found in ancient texts such as Antiphon's *Tetralogies* 'take place against a background of magical belief … they are not themselves concerned with magic … and they are not stupid'. We can understand these arguments and acknowledge their validity even if we lack the same magical 'background'. Similarly to Williams, I am not interested to argue in these pages that the OP should be read as a kind of myth or means to re-establish a 'background of magical belief'. Other, more 'merely' structural readings seem both preferable and wholly sufficient to understand the OP's function in Rawls's theory.
[10] Ibid., p. 55.
[11] Ibid.

Two kinds of fiction do this work in different ways. One is 'dense fiction', like Dickens's *Bleak House*, which 'provide[s] a depth of characterization and social background which gives substance to the moral situation and brings it nearer to everyday experience'. The other is 'stark fiction', like the tragedies of Sophocles, which direct both their 'style and structure … in a concentrated way to displaying the operations of chance and necessity'[12]. Williams' interest and ours is the latter. Sophocles' *The Women of Trachis* is cited as a paradigm of stark fiction: 'All the force of the play', Williams writes, 'is directed to leaving in the starkest relief its extreme, undeserved, and uncompensated suffering', an effect 'achieved by some complex adjustments' in its portrayal of the characters and their circumstances. And this, he says, is no mere 'unwelcome reminder of cosmic awfulness'. Rather, he argues following Nietzsche, Sophocles' play, like other tragedies, 'enables us to contemplate [this awfulness] in honesty without being crushed'[13] by it. The play's 'fictional horrors [are laid] before us in a way that elicits attitudes we cannot take toward real horrors'. No 'profound philosophical formulation'[14] can do this sort of work.

So Williams ends his essay. Surely it is hard, at first, to see the OP anywhere on this horizon. Yet as Williams suggests, 'we should not suppose … that the operations of stark fiction are always the same', so let us look more closely. The defining feature of stark fiction as Williams develops it seems to be that the 'style and structure' of such works is 'directed in a concentrated way' to displaying 'operations' not usually displayed, i.e. often hidden from view or experienced half-articulately. The operations of interest to Williams are 'chance and necessity', but there seems to be no reason why stark fiction couldn't display other operations. If this is right, we might read the notion more broadly and, I think, do so without watering down the distinction with dense fiction. For a stark fiction which displayed, say, our moral intuitions –

[12] Ibid., p. 56.
[13] Ibid., p. 58.
[14] Ibid., p. 59.

that '[laid] them before us' as no philosophical argument can do – would still not, as dense fiction does, bring these intuitions *closer* to our 'everyday experience'. The fiction would still be strange, undomesticated. Things not often expressed – tacit intuitions usually obscured by bias and circumstance – would be 'starkly' exposed. Thanks to some 'complex adjustments', the fiction would still hold us at a *distance*, not draw us closer, and precisely this distance would enable us to 'contemplate … in honesty' that which it displays, here our moral intuitions.

So I propose to extend Williams' notion of stark fiction in two ways, both of which I take to be in keeping with his own account. First, that stark fictions broadly conceived display things not usually displayed, not only chance and necessity but in general any such hidden 'operations'; and second, that by their stark character such fictions preserve a distance between these operations and our 'everyday experience', allowing us to contemplate them without prejudice and even, perhaps, to extend our 'ethical experience'. Such fictions 'do things that direct statement cannot do', and, though they may be employed in tandem with philosophical modes of reflection as Williams suggests, cannot in the last instance be 'formalized' without loss. Their dramatic effect – their confrontation with our person, both intellectual and affective – is irreducible, though not for that reason 'mythical' or 'magical'.

§2 – Rawls on the original position

Rawls distinguishes two parts of justice as fairness, his contract theory of justice: (1) 'an interpretation of the initial situation [the OP] and the problem of choice posed there', and (2) 'a set of principles which, it is argued, would be agreed to'[15]. In what follows I focus exclusively on the first of these. Rawls himself suggests the two are separable – '[o]ne may accept the first part

[15] TJ, p. 14.

of the theory (or some variant thereof), but not the other, and conversely'[16] – and such a focus will serve our present concern with the nature and function of the OP itself.

Presentation in A Theory of Justice[17]

Justice as fairness, Rawls's theory of justice, 'generalizes and carries to a higher level of abstraction the traditional conception of the social contract' (3). Any contract theory will start from some 'initial situation', and Rawls's 'original position of equality' (11) is just one interpretation of this situation. In this sense the OP 'corresponds' to (or later, 'generalizes'[18]) older state of nature views, and like these is entirely hypothetical and nonhistorical. Other interpretations of the initial situation are possible, Rawls acknowledges; like the OP, these others resemble social theoretical models (103) in seeking to enable a 'free play' of 'deliberative rationality' (496) within stipulated theoretical limits to yield a definite result. Yet most other possible interpretations, Rawls maintains, are 'irrelevant from a moral point of view' (109) – by incorporating morally irrelevant premises, they yield morally irrelevant conclusions. What makes Rawls's interpretation unique is his claim that, by carefully selecting and adjusting the conditions of the imagined initial situation, he succeeds in excluding all morally irrelevant premises. And so, the OP yields conclusions relevant 'from a moral point of view'.

How is this accomplished? To start with, what *kind* of conditions need to be selected? The conditions should be, Rawls says, 'commonly shared presumptions':

[16] Ibid.
[17] In this part I cite references to TJ in the text parenthetically. Citations from other of Rawls's works are footnoted.
[18] JF, p. 16: '[T]he original position generalizes the familiar idea of the social contract'.

> 'One argues from widely accepted but weak premises to more specific conclusions. Each of the presumptions should by itself be natural and plausible; some of them may seem innocuous and trivial' (16).

The presumptions should be 'intuitive' – 'there is an appeal to intuition', Rawls says, 'at the basis of the theory of justice' (108). The reasoning by which the OP is developed and subsequently worked on and improved must be, Rawls maintains, 'intuitive throughout' (105). No 'Cartesian' appeal is made, in the *setup* of the OP, to 'self-evident' principles from which conditions are deductively elaborated (506). Yet Rawls's theory is no mere intuitionism, thanks, precisely, to the work the OP helps us to do. For as an 'expository device', the OP holds together in one scheme all the conditions we intuitively regard as morally relevant, and *then* (what is crucial) 'helps us to extract their consequences' (19). Justice as fairness does not 'appeal' to intuition *simpliciter*; it merely takes our intuitions as a starting point, then on their basis goes further to construct a 'device', the OP, that can transform these vague subjective intuitions – transform them *for us* – into definite objective principles.

This intuitive basis of Rawls's theory is also the source of its justification, as becomes clear when one considers an obvious possible objection: Why should we real persons in the real world honor a 'hypothetical' decision reached in an ideal situation? What justifies the principles yielded by this fiction?[19] Rawls anticipates this concern in TJ: we should 'take an interest' in these principles, he writes there, because 'the conditions embodied in the description of the original position are ones that we do in fact accept. Or if we do not, then perhaps we can be persuaded to do so by philosophical reflection' (19). There must be a 'going back and forth', he says, between our intuitive judgments and the conditions we set upon the initial situation until, perhaps after many complex adjustments, the two are brought into a 'reflective equilibrium' (18).

[19] This objection was influentially articulated by Ronald Dworkin in his article 'The Original Position' (*The University of Chicago Law Review*, Vol. 40, No. 3 [Spring, 1973], pp. 500-533).

One may be sure that Rawls himself had reached such an equilibrium by the time he published TJ in 1971 – enough, at least, to offer his account of the OP. We readers are enjoined to 'enter' the OP ourselves and undertake the same process – to pose our own 'objections and replies' say, in the best modern sense, to the fruits of Rawls's meditations. Rawls claims no more than to 'have done what [he] can' (18). This process of reflection is never complete, although Rawls is convinced, pending whatever reasoned objections we might offer to sway him, that the OP as he describes it is '*the* appropriate initial status quo' (11) and the only platform able to yield *for us* the right conception of justice.

This very democratic hope that all free and equal persons might wwork together to articulate the conditions of the initial situation and consequently determine what principles it would yield goes some distance toward accounting for why Rawls retained the OP as a kind of dramatic situation or portrayal rather than making a direct formal argument for his conception of justice. For he *could* have made such an argument, a point he highlights both in TJ and again in his last book, *Justice as Fairness*[20]. In TJ he writes:

> 'If necessary, the argument to this result could be set out more formally. I shall, however, speak throughout in terms of the notion of the original position. It is more economical and suggestive, and brings out certain essential features that otherwise one might easily overlook' (120).

The meaning of 'suggestive' here may seem as obscure as an earlier claim Rawls makes, namely that the OP serves to 'make *vivid* to ourselves the restrictions that it seems reasonable to impose on arguments for principles of justice' (16). We might better understand both, however – that the OP is suggestive and vivid – by considering the meaning of Rawls's other claim: that the OP is 'economical'. For this refers not to the theory's *elegance*, I would argue, but to its *simplicity*.

[20] JF, p. 86: 'Although the argument from the original position could be presented formally, I use the idea of the original position as a natural and vivid way to convey the kind of reasoning the parties engage in'.

That the circumstances and conditions of the OP should be 'simple' is a frequent concern for Rawls. Examples of this in TJ are numerous: the conceptions of justice from which the parties are to choose should be expressed in a 'reasonably simple way' (108). They are to be presented as a list, like a 'menu' (83), a limitation itself designed to streamline the situation and make a definite choice possible. Similarly, one of the 'constraints' Rawls sets upon the parties' deliberations (constraining what principles of justice they may finally select) is 'universality' – a condition, closely connected with another constraint of 'publicity', that all 'moral persons' must be able to '*understand* these principles and *use them* in [their] deliberations'. This, says Rawls, 'imposes an upper bound of sorts on how complex [the principles selected] can be, and on the kinds and number of distinctions they draw' (114). Ultimately, seeing as the chosen 'conception of justice is to be the *public* basis of the terms of social cooperation' (122), we read that

> 'other things equal one conception of justice is to be preferred to another when it is founded upon markedly simpler general facts, and its choice does not depend upon elaborate calculations in the light of a vast array of theoretically defined possibilities. It is desirable that the grounds for a public conception of justice should be evident to everyone when circumstances permit.' (123).

Rawls's decision to present the OP as a kind of dramatic situation that all can understand and 'imagine' (17), and to retain this throughout his career, may be understood as just another of these moves intended to keep justice as fairness both simple and universally and publically accessible.

This concern for simplicity is reflected also in the precise conditions Rawls finally imposes upon the OP – the 'complex adjustments' that make his device yield a definite, morally relevant result. These adjustments and conditions make up the content of his stark fiction. They

may be considered briefly under three headings: the character and task of the parties, what they do and do not know, and the formal constraints on their deliberations.

The parties are free and equal persons 'similarly situated' (or later, 'symmetrically'[21]); not so much 'individuals' as 'continuing strands' (167). In character they are both rational, in the 'narrow sense … standard in economic theory', and mutually disinterested, 'conceived as not taking an interest in one another's interests' (12). Despite these qualities Rawls emphasizes that these are not the 'bare', interchangeable persons of utilitarian imagining (152), for they possess what he will later call the 'two moral powers'[22]: capacities (a) to *develop a sense of justice*, which Rawls thinks guarantees strict compliance with the principles selected, and (b) to *form and pursue a conception of the good* by executing a 'rational life plan'. Since the parties know *that* they have such a plan, but not its particulars, a 'thin theory of the good' is implied in the OP itself: the parties will aim to secure as many 'social primary goods' as possible – i.e. 'things that men are presumed to want whatever else they want' (230), including liberty, opportunity, wealth, and the social bases of self-respect – in order to ensure the successful pursuit of their life plans, whatever they may turn out to be.

Yet they will not do so to the detriment of their neighbors. This is an element of Rawls's design that sets the OP at some distance from similar models in social theory[23]. We read:

> 'The parties do not seek to confer benefits or to impose injuries on one another; they *are not moved by affection or rancor*. Nor do they try to gain relative to each other; they *are*

[21] PL, p. 305: '[T]he parties are symmetrically situated with respect to one another and they are in that sense equal'.
[22] See for example PL, p. 308.
[23] As we read in JF, p. 81: 'Our aim is to uncover a public basis for a political conception of justice, and doing this belongs to political philosophy and not social theory'.

> *not envious or vain*. Put in terms of a game, we might say: they try strive for as high an absolute score as possible [in terms of social primary goods]' (125).

Who *are* these cool, monadic parties? And how can Rawls justify these seemingly fanciful stipulations about their characters – no envy? no vanity? Later in *Justice as Fairness*, in what is more a clarification of his account in TJ than an extension, Rawls justifies his claims about this whole class of 'special psychologies'. He writes:

> 'Remember it is up to us, you and me, who are setting up justice as fairness, to describe the parties … as best suits our aims in developing a political conception of justice. Since envy, for instance, is *generally regarded* as something to be avoided and feared … it seems desirable that, if possible, the choice of principles should not be influenced by this trait. So we stipulate that the parties are not influenced by these psychologies'[24]

Since we would ('intuitively') like to think that petty biases do not determine the principles we live by, and since the OP is our own fiction – we are its authors – nothing prevents us from simply excluding these influences. Crucially however, if our fiction is to be not a fiction *merely* but rather a device able to yield objective principles for a well-ordered society, we must 'consider', Rawls says in TJ, 'whether the conception arrived at [under these stipulations] is feasible in view of the circumstances of human life' (124). If it is not – if we find that envy and vanity are somehow permanently written into ourselves and not just the products of unjust circumstances – then we must go back to the beginning and work out a new version of the OP that incorporates these traits. But Rawls, for one, thinks this will not be necessary. '[O]ur nature', he writes, 'is such as to allow the original choice to be carried through. In this sense we might say that humankind has a moral nature' (508).

[24] JF, p. 87.

So much for the parties themselves. Just as important as these characters' qualities are the conditions placed on their deliberations, i.e. what they do and do not know. They *do* know two sorts of things. First, 'whatever general facts affect the choice of the principles of justice', including all 'general laws and theories' such as the principles of economics and laws of human psychology (119)[25]. Second and more importantly, the parties know 'particular', even empirical facts. Rawls calls these particular facts taken together the 'circumstances of justice', or 'normal conditions under which human cooperation is both possible and necessary' (109). Two kinds of circumstances are distinguished. First, *objective* circumstances: namely that we are individuals of roughly equivalent powers, and that we exist together, vulnerable to each other, each pursuing his or her own plans under material conditions of moderate scarcity. Alongside these are *subjective* circumstances: that since we pursue different conceptions of the good and make claims on the same material resources, our respective plans shall often conflict (a condition described in Rawls's later works as 'the fact of reasonable pluralism'[26]). What's more, we find ourselves marked by 'shortcomings of knowledge, thought, and judgment' as well as 'moral faults', even if these latter are often 'simply part of [our] natural situation' (110). Taken together these circumstances 'set the stage', as Rawls puts it, 'for questions of justice' (112).

[25] These facts are assumed to be true and correct (TJ, p. 481), but, as always for Rawls, are judged so only from our own point of view, not from that of a 'transcendent being' or a 'place beyond the world' (TJ, p. 514). We ourselves determine what shall count as a fact in the only way we can: by relying upon 'current knowledge as recognized by common sense and the existing scientific consensus' (TJ, p. 480). In this sense these facts are facts just for 'you and me, here and now' – they are just elements of our own 'best account', say, of how things are. And yet, though lacking a kind of Cartesian deductive certainty to which Rawls never aspires in his setup of the OP, they crucially are not for that reason so many mere *façons de parler*. For us they really are the facts of our situation, and we cannot do otherwise, or more, than to rely upon them in determining principles of justice.

[26] In his discussion of the circumstances of justice in JF (p. 84), Rawls writes: 'We take this pluralism [i.e., reasonable pluralism] to be a permanent feature of a democratic society, and view it as characterizing what we may call the subjective circumstances of justice'.

Thus the curtain rises, so to say, on the parties' deliberations. Right away each finds him or herself shrouded in a 'veil of ignorance' – for there are many things they do *not* know. Rawls highlights four: their class or standing in society; their natural assets and abilities, such as intelligence and strength; the particulars of their own conception of the good and special features of their psychological disposition; and the particular circumstances of their society, i.e. its political or economic situation and the 'level of civilization and culture it has been able to achieve' (118). Even 'the historical record is closed to them' (162).

Rawls's casting of this veil seems motivated by a fourfold purpose. First, it serves to 'nullify the effects of specific contingencies which put men at odds and tempt them to exploit social and natural circumstances to their own advantage' (118), so ensuring that only morally relevant premises are considered. Second, it makes possible a unanimous agreement among the parties. Because 'everyone is equally rational and similarly situated' we can be sure that 'each [party] is convinced by the same arguments. … If anyone after due reflection prefers a conception of justice to another, then they all do' (120). In this way, he maintains, 'a genuine reconciliation of interests' (122) can be achieved. Third, the veil ensures 'not only that the information available is [morally] relevant, but that it is always the same' (120). And of course: when the *same* parties consider the *same* principles with the *same* information, as they always do – and when the same purely 'deductive' form of deliberation prevails within these reasonable conditions[27], as it always does – one can be certain that 'the same principles are always chosen' (ibid.). Fourth and finally, the veil ensures simplicity. As Rawls puts it,

[27] In JF (p. 82) we read: 'We should like the argument from the original position to be, so far as possible, a deductive one, even if the reasoning we actually give falls short of this standard'. Rawls adds a footnote to this comment (n. 3), citing TJ pp. 104f, where we read: 'The argument [from the OP] aims eventually to be strictly deductive'. It is important to understand that, although the setup of the OP proceeds from an 'intuitive basis', the deliberations of the parties are by no means 'intuitive', but strictly rational 'in the narrow sense, standard in economic theory' (TJ, p. 12). One does well to juxtapose these expressions of

> '[w]ithout these limitations on knowledge the bargaining problem of the original position would be hopelessly complicated. Even if theoretically a solution were to exist, we would not … be able to determine it' (121).

The veil keeps simple, one might say, both *the bargaining* of the parties, i.e. the technical terms in which their deliberations are conducted, and also *the problem* itself, i.e. the task defined by the OP that Rawls thinks should be intelligible to all persons ('if circumstances permit') whatever their share in the natural asset of intelligence. Without the veil of ignorance 'the bargaining problem' of the OP would be no less complex than deliberations on the floor of Congress, and one imagines no more generally intelligible.

And so the parties begin their deliberations, proceeding rationally to select a conception of justice from among those 'given [on] a short list of traditional conceptions' (106). The chosen conception will be the one that 'ranks' highest among those given. In making their ranking, the parties find themselves constrained in several ways by 'the concept of right', mentioned above – this being the last of Rawls's 'complex adjustments'. Five such constraints are listed in TJ, of which only three are retained in his later work[28]. These three are the most relevant to our

---

Rawls's deductive ambitions with a qualification he issues on p. 133 of JF: the ideal of a fully deductive approach, he notes there, means only that 'all the necessary premises for the argument from the original position … are included in the description we gave of it' – or rather, say, that a 'free play' of 'deliberative rationality' has truly been enabled within the given limits (TJ, p. 496). Yet this deductive ideal will not be fully attainable, he continues (in JF, pp. 133-4), both because 'there are indefinitely many considerations that may be appealed to in the original position', and because 'the balance of reasons itself rests on judgment' – i.e. we *judge* which of the conceptions ranks most highly, and, though this is 'informed and guided by deductive reasoning', such an individual judgment can never be purely analytic and so never purely deductive.

[28] Cf. Rawls's mature treatment of the constraints of the concept of right in JF, p. 86. Generality, university, and publicity are all noted, but the others named in TJ pp. 113f. are left out: namely that the principles should be *systematic*, such as to enable an 'ordering' of 'conflicting claims', and *final*, being able to serve as 'the final court of appeal in practical reasoning … [such that] reasoning successfully from these principles is conclusive' (TJ, p. 116). That these two in particular should have been omitted from

purposes: that the principles selected should be *general*, not dependent upon 'proper names' (113) and so able to serve as 'a public charter of a well-ordered society in perpetuity' (114); *universal*, 'hold[ing] for everyone in virtue of their being moral persons'; and *public*, such that 'everyone will know about these principles' and be aware of their 'universal acceptance' (115). Rawls emphasizes that these constraints are not derived from a formal 'analysis' of the concept of right – 'many constraints', he says, 'can reasonably be associated with [that] concept' (112), and these are just some of those possible. That these are however *the* right constraints *for us* to impose in the OP will become clear, he says, from 'the soundness of the [resulting] theory itself' (113). If this theory proves less than 'feasible' for us, we must reassess our 'reflective equilibrium' – reconsider both the conditions of the OP and our own intuitive judgments – and make the necessary revisions.

Such is Rawls's account of the OP as set out in TJ. Although all these main elements remained in place throughout his career, a new description of it would come to play an important role in his theory.

Rawls's new description: The original position as a 'device of representation'

The single major change in Rawls's later account of the OP is his description of it as a 'device of representation'. This term does not appear in TJ, although Rawls does there suggest that the OP could serve as 'a useful analytic *device*'[29], and I've employed this language already above. The new representational character of this 'device' image needs fleshing out, however. It lends crucial support to my proposed reading.

Rawls's later presentation is not surprising given the political ('not metaphysical') turn in his writing from PL onward.

[29] TJ, p. 65.

The term 'device of representation' first occurs in PL, Rawls's first major work after TJ, in lecture one. We read: '[T]he original position is simply a device of representation … [E]ach of [the parties] is *responsible for* the essential interests of a free and equal citizen …'[30]. To understand this new description, one must consider Rawls's new distinction, a few pages later, of 'three points of view' relevant to the OP: '[1] that of the parties in the original position, [2] that of citizens in a well-ordered society, and finally, [3] that of ourselves – of you and me who are *elaborating* justice as fairness and *examining* it as a political conception of justice'[31]. The first two of these he says 'belong to the conception of justice as fairness'; the third 'is that from which justice as fairness is to be assessed'[32] – assessed, but also, as noted above, 'elaborated' or authored. Nowhere in TJ does Rawls distinguish these three points of view explicitly.

This new distinction signals two important changes in Rawls's account of the OP. First, the parties in the OP are made into 'representatives'[33] – not representatives of ourselves, but of 'citizens in a well-ordered society'. This move creates a new kind of *distance* between the parties and the citizens, which two examples may be taken to illustrate. In TJ, the parties lacked knowledge of their *own* conception of the good, and were disinterested in *each other*. Here in PL, however, both these conditions undergo a shift: now, the parties lack knowledge of the 'conception of the good *of the persons they represent*'[34], and 'take no direct interest in the interests *of persons represented by other parties*'[35]. This new remove implied by the language of 'representatives' and 'constituents' is emphasized in other ways as well: e.g., we are told that the parties 'act as trustees or guardians … [aiming to] secure the fundamental interests *of those they*

[30] PL, p. 25.
[31] PL, p. 28. Emphasis added.
[32] Ibid.
[33] PL, p. 25.
[34] JF, p. 88.
[35] JF, p. 85.

represent'[36] – that is, not their *own* interests. Also crucial is Rawls's introduction of the distinction between the merely 'rational autonomy' of the parties (which is not an 'ideal', but just 'a way to model the idea of the rational' in the OP) and the 'full autonomy' of citizens (which *is* 'a political ideal')[37]. In all these ways, the parties are set at a distance from the citizens they represent. They are placed in the citizens' service.

A second change to Rawls's account follows from this: namely, that a distance analogous to that between parties and citizens is opened up between, on the one hand, *the whole of Rawls's theory*, justice as fairness – i.e. *both* the dramatic device of the OP (the theory's 'first part', as distinguished above) *and*, as its deductive consequence, the principles and citizens it yields (its second part) – and, on the other hand, ourselves, we who 'elaborate' and 'assess' the OP from Rawls's third 'point of view'. Here in PL, we are set apart from the (theoretically internal) dynamic between the parties and citizens. We may 'enter' this dynamic for a time, when we engage with Rawls's theory, yet remain decisively outside it. As Rawls's writes in *The Law of Peoples,* '[w]hat is modeled [in the OP] is a *relation* … of the parties representing citizens'[38] – not of the parties representing either 'us' or themselves. Rawls's explicit distinction of the parties' and citizens' points of view – the only two that 'belong to … justice as fairness' – serves to make the OP more clearly self-enclosed than it may have seemed in TJ, emphasizing that the theory is a no more than a construction (say, a fiction) of which we 'ourselves', outside it, are the authors and assessors.

[36] JF, pp. 84-5. Cf. PL, p. 307: Among 'the considerations that move the parties in the original position,' Rawls writes, 'their overall aim is to fulfill their responsibility and to do the best they can to advance the determinate good of the persons they represent'.
[37] PL, p. 28.
[38] LP, p. 30, n. 32. Emphasis in original.

Rawls admits that his presentation of the OP in TJ may have produced 'misunderstandings'[39] about justice as fairness, notably a sense that the theory entails a conception of persons as, 'in their essential nature … independent of and prior to their contingent attributes'[40]. Such has indeed been a common complaint. Yet this, Rawls continues, is an 'illusion … caused by not seeing the original position as a device of representation'[41]; caused, that is, by not distinguishing adequately between our own point of view on the one hand, and those of the parties and citizens on the other; by not taking sufficient distance from the construction as a whole. Rawls goes on to dismiss this 'illusion' in significant terms:

> 'When … we simulate being in the original position, our reasoning no more commits us to a particular metaphysical doctrine about the nature of the self than our acting a part in a play, say of Macbeth or Lady Macbeth, commits us to thinking that we are really a king or a queen engaged in a desperate struggle for political power. Much the same holds for role playing generally'[42].

One may thus read the OP, I would argue, as a unified dramatic portrayal of justice as fairness, and so see it as including *both* the parties and, by implication insofar as these parties have become 'representatives', *also* the citizens to which they are beholden. For who are these citizens, in concept, but the 'products' of the OP – the 'resolution' of its dramatic tension? Having been formed in a well-ordered society governed by the principles of justice, the very existence of these citizens determines that, as a drama, the OP shall be not a tragedy like *Macbeth* or *The Women of Trachis*, but a comedy.

By his distinctions, then, *first* between parties and citizens, and *then* between both and ourselves – both of these proceeding from his new description of the OP as a 'device of

[39] PL, p. 28.
[40] PL, p. 27. In a note (n. 29), Rawls directs the reader to 'the important work of Michael Sandel, *Liberalism and the Limits of Justice* [1982]'.
[41] Ibid.
[42] Ibid.

representation' – Rawls brings his 'idea' of the OP full circle, filling out and completing its dramatic structure, and so empowering us ('you and me, here and now') both to apprehend it as spectators and then, on this basis, to work to achieve the well-ordered society it promises.

§3 – The original position: structure and function

Let us return to the two questions posed above: what kind of idea is the OP, and what work does it do? The first I take to be a question about the idea's *structure*, and so shall begin by defending my reading of it as a stark fiction. The second I take to be concerned with its *function*, and so shall conclude with a defense of my main and more fundamental claim: that the OP, read as a stark fiction, serves as a 'mediating idea' in justice as fairness by means of which heteronomous individuals may be transformed into fully autonomous citizens.

I take the easier point first: the OP is a 'fiction'. Rawls implies as much, at least in a thin sense, by his insistence on the idea's nonhistorical character and description of it as a 'thought experiment'. In JF this latter description is given a more thick, dramatic sense: the parties, we read, are 'merely … *characters* who have a part *in the play* of our thought experiment'[43]. Like *Macbeth*, or say the prisoner's dilemma[44], the OP portrays a situation – the making of an agreement – which never has and indeed never *could* occur in the real world; as Rawls adds in JF, 'even if it could [occur], that would make no difference'[45]. For consider another example: what difference would it make if we were to learn that Cervantes had modeled the character of Don Quixote upon that of a real historical figure in sixteenth century Spain? Such a revelation could even *diminish* the power of Cervantes's fiction by mitigating our ability to suspend disbelief in the story, making us wonder at every moment whether Cervantes 'got it right' – is

[43] JF, p. 83. Emphasis mine.
[44] TJ, p. 238.
[45] JF, p. 16.

this really the 'way things were' for the real Don Quixote? Precisely because we recognize Cervantes's work as a fiction, we can enter wholeheartedly into its narrative with fewer concerns about its *accuracy* with respect to history, and more for its *adequacy* with respect to *our own* (moral) experience. The seemingly perennial capacity of *Don Quixote* to prove adequate in this latter sense is why we call it a great work. In the same way, Rawls makes clear in both PL and JF that justice as fairness should be regarded as our own 'undertaking'[46]. *We ourselves* must decide what to regard as 'a fair system of cooperation between free and equal citizens', for this cannot be 'given' by an outside authority – be it 'God's law' or 'an independent moral order'[47]. As with Cervantes's novel, our concern should be not whether the situation portrayed in the OP is historically *accurate* – this would do nothing to strengthen its justification – but whether it is *adequate* to our own intuitions and so able to yield principles that could 'feasibly' govern our real society.

Yet *Don Quixote* is probably a dense fiction, and my contention is that the OP is a stark one. If this is right, then the OP can do certain things that a fiction like *Don Quixote* cannot do, just as *The Women of Trachis* serves us differently than *Bleak House*. We've seen above what these things are, according to my reading of Williams: stark fictions display real 'operations' not usually displayed, yet preserve a distance between ourselves and the spectacle taken whole, thus preserving its strangeness and so its ability to challenge the supposed adequacy of our everyday point of view. Rawls's 'complex adjustments' to the OP, combined with his later description of it as a device of representation, seem to achieve these effects. As Rawls puts it in PL, the OP 'helps us work out what we now think … [by offering] a *clear and uncluttered* view of what justice

[46] PL, p. 22.
[47] Ibid.

requires'[48] – a view we win only by imposing the veil of ignorance. Similarly, in JF we read that the OP 'provides a way to keep track of our assumptions … [and] *brings out the combined force* of [these] assumptions by uniting them into one surveyable idea that enables us to see their implications more clearly'[49].

So much for the OP's structure. Let us consider its function, which I've suggested is the formation of citizens. This claim is best elaborated by considering *how* the OP accomplishes this task, and so I begin with this. Although I've argued that in form the OP is both stark and simple, this should not of course be taken to mean that it lacks intricacy, for its stark and simple form is achieved not only by the very complex adjustments reviewed above, but also, I argue, by the employment of a sophisticated dramatic device: *a play within a play*. Let me explain.

To claim that the OP may be read as a kind of play is perhaps not controversial, being no more than a benign redescription of the idea that one can foist even upon Rawls's account in TJ without much trouble. Rawls himself, as we've seen, seems to suggests this. Yet Rawls's new distinction in PL of three points of view – of the parties, the citizens, and ourselves – may be drawn upon to support a stronger claim. Let's assume for a moment that my suggestion above concerning the distance between ourselves on the one hand and the parties and citizens on the other is correct. If so, then we ourselves are the spectators of a kind of play – that's the first, benign level. But there is another: for *within* this play, some of the characters are *themselves* watching a play – i.e. the citizens themselves are imagining the situation of the parties in the OP (as we ourselves have defined it) and assessing and re-elaborating that situation, just as we are. In this sense, the parties in the OP (1st point of view) function as a common dramatic and

[48] PL, p. 26.
[49] JF, p. 81.

mediating point of reference for the other two points of view (of the citizens and ourselves). Our relation to the parties is in this sense different than our relation to the citizens, though apprehended from the same remove: we represent ourselves *as* the parties, but *do not desire* their merely rational autonomy; we represent ourselves *as* the citizens, and precisely *do desire* their full autonomy. In TJ we had just one role to play, one point of view to adopt: that of the parties. This enabled a kind of rational assessment of our everyday situation and made possible, again merely 'rationally', a selection of the two principles. There was here indeed a kind of dialectic of identification and distancing, but only in the first degree. One is not surprised that this first account gave rise to the sorts of metaphysical 'illusions' to which Rawls responds in PL.

Yet once the OP is read as a device of representation, and once the point of view of the citizen is explicitly distinguished and incorporated as part of the idea's unified spectacle, a second and more transformative degree of this dialectic becomes possible. Rather than identifying with the parties, we can identify with the *citizens*; we can imagine ourselves *as* citizens imagining themselves *as* the parties, and so adopt a kind of *double*-remove. We thus become like spectators of the play within a play in *Hamlet*, and like Hamlet himself are enabled not only better to *assess* our situation by viewing it through this spectacle – TJ enabled that well enough – but, more than this, are motivated to *act* and to allow ourselves to be transformed by our vision of the citizens' full autonomy, say by our anticipatory experience of it *as if* it were our own, from an aesthetic remove. So attractive is this full autonomy, both intellectually and affectively, that we may desire to *close* this remove, to identify as citizens in our real lives, and so leave Rawls's third, everyday point of view behind. We may want, that is, to trade our present heteronomy for the 'political ideal' of a citizens' full autonomy, and are *made* to want this by the

dramatic device of the OP, for, by working through this device with a sufficient suspension of disbelief, it is *as if* we had already lived it.

To summarize: beginning from our everyday, heteronomous point of view (Rawls's third), we elaborate the OP as a device of representation with all its complex adjustments. In it, we model ourselves both as rationally autonomous agents (the parties) and as fully autonomous 'free and equal persons' (the citizens). By *identifying* with the citizens through a kind of role-playing – by recognizing ourselves in them, one might say – we find ourselves motivated to undertake with greater energy and conviction the (merely) rational exercise that is the condition of the possibility of these citizens' happy state – a state we desire. This *desire*, enabled by the OP's dramatic form and which cannot arise merely from our relation to the parties, drives the entire process, thus serving as an engine 'under the hood', as I've said, of justice as fairness. This dynamic is not apparent unless one reads the OP as a stark fiction.

The OP is, then, a 'mediating idea' in a more robust sense than Rawls himself seems to have intended. Not only does it mediate a transformation of our often contradictory subjective moral intuitions into publically affirmable objective principles by holding these together in a 'clear and uncluttered' spectacle; it also, more deeply one might say, mediates a transformation of ourselves as heteronomous individuals into fully autonomous citizens.

I conclude with an important note about the very democratic nature of this process. For on the reading I've been advocating, and contrary to what Rawls's critics often maintain, justice as fairness is not merely a kind of academic exercise best undertaken by professors and graduate students – i.e. those with enough leisure and philological acuity to work through the whole corpus of Rawls's work, 'chapter and verse'. For although this kind of careful attention to the particulars of Rawls's theory is of course necessary, it is not ultimately what makes the theory *work*. The

theory ‘works’, I would say – i.e. becomes able to serve as a kind of platform from which *everyone* can identify as a citizen and undertake the project of building a more just society – only because it appeals to our whole person, intellectually *and* affectively, through the ‘stark’ and ‘simple’ dramatic device of the OP. I began these pages with the suggestion that ‘no idea is more closely associated with the work of John Rawls than that of the original position’. Why is this so? Not, certainly, because the idea is *theoretically* crucial to his system – as we’ve seen, Rawls himself says he could do without it – and not because the parties themselves are in any way compelling. Rather, the OP draws so much public attention – both praise and detraction – *because* it appeals to our imagination and desire, in such a way that even those only vaguely familiar with Rawls’s writings can ‘enter’ the situation it depicts and ‘work through’ it – or against it, as they like. Only in this deliberative and participatory manner can justice as fairness hope to ‘generate its own support’ (119), and so transform our societies and ourselves.

Bibliography

Rawls, John. (1993) *Political Liberalism*. New York: Columbia UP.
----------. (1999) *A Theory of Justice*. Rev. ed. Cambridge: Belknap Press.
----------. (1999) *The Law of Peoples*. Cambridge: Harvard UP.
----------. (2001) *Justice as Fairness: A Restatement*. Cambridge: Belknap Press.
Williams, Bernard. (2007) '*The Women of Trachis*: Fictions, Pessimisms, Ethics'. In *The Sense of the Past: Essays in the history of Philosophy*. Princeton: Princeton UP, pp. 49-59.