Original Position Models, Trade-offs and Continuity

STEVEN DASKAL
Northern Illinois University

John Harsanyi has offered an argument grounded in Bayesian decision theory that purports to show that John Rawls's original position analysis leads directly to utilitarian conclusions. After explaining why a prominent Rawlsian line of response to Harsanyi's argument fails, I argue that a seemingly innocuous Bayesian rationality assumption, the continuity axiom, is at the heart of a fundamental disagreement between Harsanyi and Rawls. The most natural way for a Rawlsian to respond to Harsanyi's line of analysis, I argue, is to reject continuity. I then argue that this Rawlsian response fails as a defence of the difference principle, and I raise some concerns about whether it makes sense to posit the discontinuities needed to support the other elements of Rawls's view, although I suggest that Rawls may be able to invoke discontinuity to vindicate part of his first principle of justice.

John Rawls's social contract analysis of justice famously involves the use of a hypothetical original position from which we are to imagine individuals choosing principles of justice. Some commentators downplay the significance of his original position argument and instead focus on other elements of Rawls's view, such as his arguments involving reciprocity, stability and the strains of commitment. I do not deny the significance of those arguments, but I think they are best understood, and most effective, when situated in conjunction with Rawls's original position model, which undeniably plays a key role in Rawls's own presentation of his view, both in *A Theory of Justice* and later in *Justice as Fairness*.[1] Moreover, as I will explain, the original position strikes me as capturing an important intuition about justice and morality. What I find less clear is whether an original position argument, properly developed, yields the substantive conclusions endorsed by Rawls. That is what I will be exploring in this article. In particular, I will be considering a challenge to Rawls, raised initially by John Harsanyi and R. M. Hare, and recently endorsed by Allan Gibbard, according to which an original position analysis leads inexorably to utilitarian conclusions.

[1] John Rawls, *A Theory of Justice*, rev. edn. (Cambridge, MA, 1999) and *Justice as Fairness: A Restatement* (Cambridge, MA, 2001). In this article I will focus primarily on Rawls's view as presented in *Justice as Fairness*, which he describes as rectifying some mistakes from *Theory* and unifying the view from *Theory* with that developed in many of his subsequent essays. See Rawls, *Justice as Fairness*, pp. xv–xviii.

© Cambridge University Press 2015
Utilitas Vol. 28, No. 3, September 2016
doi:10.1017/S0953820815000278

CrossMark
https://doi.org/10.1017/S0953820815000278 Published online by Cambridge University Press

One difficulty in assessing this challenge is that it has often been formulated, particularly by Harsanyi, in a style of argument so different from Rawls's that it can be difficult to identify precisely where the disagreements lie. Moreover, Harsanyi's formulation exposes the argument to potential objections that, although important for assessing Harsanyi's argument, may not capture the most fundamental points of disagreement between Harsanyi and Rawls. One aim of this article is therefore to bring Harsanyi's line of argument into more direct contact with Rawls's analysis and identify what I take to be the key points of contention, hopefully without doing excessive violence to either of their views. Once that has been done, I will be able to explore the extent to which Rawls is able to rebut the objection. My ultimate conclusion will be that in the face of this challenge Rawls gets substantially less than he wants, but perhaps somewhat more than his critics acknowledge.

I begin, in section I, with an explanation of Rawls's original position analysis and the utilitarian challenge, emphasizing Harsanyi's formulation of the challenge which is grounded in Bayesian decision theory. In section II, I consider and reject a possible Rawlsian response that is focused on precluding the possibility of probability judgements, and in doing so further clarify the nature of the challenge. In section III, I introduce what I believe turns out to be a more promising line of response on Rawls's behalf, initially expressed in terms of intolerable outcomes. In sections IV and V, I refine this Rawlsian response by recasting it in terms of value discontinuity, and argue that a seemingly innocuous Bayesian rationality assumption, the continuity axiom, is at the heart of a fundamental disagreement between Harsanyi and Rawls. The most natural way for a Rawlsian to respond to Harsanyi's line of analysis, I argue, is to reject continuity. In section VI, I explore the idea of value discontinuity, and then, in section VII, I evaluate the discontinuity response as a strategy for defending Rawls from arguments like Harsanyi's. I argue that this Rawlsian response fails as a defence of the difference principle, and I raise some concerns about whether it makes sense to posit the discontinuities needed to support the other elements of Rawls's view, but I suggest that Rawls may be able to invoke discontinuity to vindicate part of his first principle of justice.

I. RAWLS'S CONCEPTION OF THE ORIGINAL POSITION AND THE UTILITARIAN CHALLENGE

For Rawls, the fundamental idea guiding the development of an adequate theory of justice is that society ought to be structured as a fair system of social cooperation. As he sees it, a primary obstacle to developing principles of justice that succeed in this regard is that

https://doi.org/10.1017/S0953820815000278 Published online by Cambridge University Press

actual agreements between actual people often reflect self-interest and relative bargaining power rather than fairness. He introduces the original position as way of overcoming this obstacle and modelling agreements that accurately reflect fairness. The idea is that we are to imagine discussants in a hypothetical situation, and we are to construct the situation in such a way that any agreement reached in that situation is guaranteed to capture the demands of fairness. This hypothetical situation is what he calls the original position, and his original position argument for his preferred principles of justice involves first defending a specific characterization of the original position and then arguing that those principles would be chosen by hypothetical individuals deliberating within the original position so characterized.

The key to Rawls's characterization of the original position is that it is a situation in which one must deliberate on the basis of a restricted set of information. In particular, Rawls sets up the original position so that parties in the original position are forced to choose principles of justice for a society without knowing their own roles or positions within that society.[2] They must deliberate behind what he calls a veil of ignorance that prevents them from knowing self-pertinent information that would potentially lead them to prefer, and bargain for, principles of justice that benefit themselves specifically. Among other things, the veil precludes knowledge of one's race, gender, religious affiliation, and natural talents and abilities. Rawls's idea is that by excluding this self-identifying information the veil of ignorance eliminates the possibility of self-interested bias, and thereby ensures that deliberation and choice within the original position gives appropriate weight to the interests of all members of society. Whatever principles of justice they would choose, or agree to, under these conditions will therefore be ones that are appropriate for governing a society conceived as a system of fair cooperation.

Given this characterization of the original position, Rawls then argues that individuals deliberating within the original position would choose the following set of principles of justice. To begin with, they would agree upon a first principle guaranteeing equal basic liberties to all. In addition, they would agree upon a second principle governing social and economic inequality in such a way that (a) relative advantages and disadvantages are connected to positions that are genuinely open to all, and (b) these relative advantages and

[2] In Rawls, *Theory*, the discussants are imagined as choosing principles for a society in which they will (or do) live, whereas in Rawls, *Justice as Fairness*, they are redescribed as trustees, so that they are choosing principles for a society on behalf of one who will live (or does live) in that society. This distinction does not matter for my analysis.

https://doi.org/10.1017/S0953820815000278 Published online by Cambridge University Press

disadvantages are structured in a way that is of greatest benefit to the least advantaged. Part (a) of the second principle ensures what Rawls calls 'fair equality of opportunity' and part (b) imposes what he calls 'the difference principle'. I will discuss Rawls's principles of justice at greater length in section VII, but for now the important point is that Rawls thinks of them as constituting a stark contrast to a standard utilitarian principle that simply dictates maximizing the total or average utility within the society. As he understands it, the principal challenge for his analysis is to show that his preferred principles are superior to utilitarian alternatives.[3]

At least within the confines of Rawls's original position argument, that superiority is manifest in the fact that discussants in the original position would choose his principles over utilitarian ones. Or so Rawls claims. The trouble is that there is a fairly simple argument that purports to show the opposite; it purports to show that deliberation within Rawls's original position leads directly to utilitarian conclusions in a relatively straightforward way. I believe Hare was the first to raise this argument as an objection to Rawls, although the line of argument itself was developed earlier, both by William Vickrey and, in greater detail, by Harsanyi.[4] As Hare sees it, Rawls is right to think that the correct principles of justice are ones that would be chosen by individuals deliberating behind a veil of ignorance that screens off the possibility of self-interested bias. But, he claims, the appropriate deliberative strategy in such a situation is to evaluate competing principles of justice as follows. For each set of principles in contention, consider the odds of occupying each role or position available in a society governed by those principles, determine how good or bad it would be to occupy each of those positions, and then perform an expected utility calculation that multiplies the desirability (or utility) of each position by the chance of

[3] Rawls is, of course, also concerned to argue against other alternatives, such as libertarian principles, particularly 'right-libertarian' ones. But, especially in *Justice as Fairness*, he frames his analysis primarily as an attempt to demonstrate that his principles are preferable to utilitarian ones, and then points out the ways in which the arguments developed for that purpose also suffice to show that his principles are superior to other contenders. The emphasis on utilitarianism as the primary alternative to his view is perhaps most evident from the fact that Rawls structures his argument as involving two 'fundamental comparisons', first between his preferred principles and an unrestricted form of utilitarianism and second between his principles and restricted forms of utilitarianism.

[4] R. M. Hare, 'Rawls' Theory of Justice', *Reading Rawls*, ed. Norman Daniels (New York, 1975), pp. 81–107; William Vickrey, 'Measuring Marginal Utility by Reactions to Risk', *Econometrica* 13 (1945), pp. 319–33. Harsanyi's earliest version of the objection is in John C. Harsanyi, 'Cardinal Utility in Welfare Economics and in the Theory of Risk-Taking', *Journal of Political Economy* 61 (1953), pp. 434–5. Harsanyi asserts that his own analysis was independent of Vickrey's in John C. Harsanyi, 'Morality and Theory of Rational Behavior', *Social Research* 44 (1977), pp. 623–56, at 634.

https://doi.org/10.1017/S0953820815000278 Published online by Cambridge University Press

occupying that position and then aggregates or averages the results. If one does not know one's role in society, the sensible thing to do is to choose whatever principle or set of principles yields the highest expected utility in this calculation. In other words, parties in Rawls's original position, according to Hare, ought to choose as utilitarians. This does not necessarily mean they will choose a principle of utility as the ideal principle of justice to govern a society: both advocates and critics of utilitarianism have argued that there can be cases in which utilitarian reasoning leads to the endorsement of non-utilitarian principles, and perhaps this is one of them.[5] But it does transform the task of defending Rawls's preferred principles of justice into the task of showing that they have a deeper utilitarian rationale, which is certainly not the task Rawls sees himself as performing.

Well before Hare raised this point as an objection to Rawls, and even before Rawls's earliest published formulation of the line of analysis that he later refined into his original position argument, Harsanyi had been developing an argument meant to show that deliberation under conditions like those found in Rawls's original position appropriately takes a utilitarian form.[6] The focal point of Harsanyi's argument is what he calls the equiprobability postulate, which dictates the following:

> If we want to decide between two alternative moral standards A and B, all we have to do is ask ourselves the question, 'Would I prefer to live in a society conforming to standard A or in a society conforming to standard B? – assuming I would not know in advance what my actual social position would be in either

[5] There is a great deal of debate and literature on this. For an example of an argument of this sort posed as an objection to utilitarianism, see Bernard Williams, 'A Critique of Utilitarianism', J. J. C. Smart and Bernard Williams, *Utilitarianism: For and Against* (Cambridge, 1973), pp. 77–118, at 134–5. For examples of arguments of this sort incorporated into a utilitarian perspective, see R. M. Hare, *Moral Thinking: Its Levels, Method and Point* (Oxford, 1981), and Peter Railton, 'Alienation, Consequentialism, and the Demands of Morality', *Philosophy and Public Affairs* 13 (1984), pp. 134–71.

[6] Rawls's earliest published formulation of the idea that developed into his original position analysis is in John Rawls, 'Justice as Fairness', *The Journal of Philosophy* 54 (1957), pp. 653–62. Harsanyi initially presents his version of an original position (or equiprobability) argument in Harsanyi, 'Cardinal Utility', and presents a related axiomatic argument in John C. Harsanyi, 'Cardinal Welfare, Individualistic Ethics, and Interpersonal Comparisons of Utility', *Journal of Political Economy* 63 (1955), pp. 309–21. He deploys the original position argument as a criticism of Rawls in John C. Harsanyi, 'Can the Maximin Principle Serve as a Basis for Morality? A Critique of John Rawls's Theory', *The American Political Science Review* 69 (1975), pp. 594–606, and develops his own view more fully and explicitly, presenting the two arguments side by side, in John C. Harsanyi, 'Morality and Theory of Rational Behavior', *Social Research* 44 (1977), pp. 623–56. For an excellent analysis of Harsanyi's arguments, focusing on the axiomatic version and developing a similar argument of his own, see John Broome, *Weighing Goods: Equality, Uncertainty and Time* (Oxford, 1991).

https://doi.org/10.1017/S0953820815000278 Published online by Cambridge University Press

society but rather would have to assume to have an equal chance of ending up in any one of the possible positions.'[7]

With this postulate in place, and what he characterizes as a few very weak Bayesian rationality assumptions, Harsanyi works through a formal proof that is intended to yield standard utilitarian conclusions. Moreover, Harsanyi views Rawls's endorsement of the principles of justice one would choose behind a veil of ignorance in the original position as committing Rawls to this equiprobability postulate. As a result, Harsanyi takes his proof to demonstrate that Rawls's original position analysis leads directly to utilitarianism.

Harsanyi's attempted proof has been the subject of considerable debate. Perhaps the most significant criticism, sometimes referred to as the standard objection to Harsanyi, involves the distinction between ordinal and cardinal rankings.[8] The fundamental problem, according to this objection, stems from Harsanyi's use of von Neumann–Morgenstern utility functions, or vNM utility functions, to represent the value, or utility, an agent assigns to possible outcomes. The trouble is that Harsanyi illegitimately supposes that representing preferences with vNM utility functions reveals a definitive cardinal ranking of those preferences. Harsanyi needs this cardinal ranking in order to make comparisons of preference strength, either intrapersonally, of the form 'A has a stronger preference for X over Y than for Y over Z', or interpersonally, of the form 'A prefers X over Y more than B prefers Y over X'.[9] What Amartya Sen, John Weymark and John Roemer point out is that the axioms of vNM expected utility theory are not sufficient to impose cardinality on preferences. And the merely ordinal ranking, which is all that Harsanyi is entitled to rely on, is too weak for the purposes of his proof.

[7] Harsanyi, 'Morality', p. 634.

[8] This is identified as the 'standard objection' to Harsanyi in John Broome, 'Can There Be a Preference-Based Utilitarianism?', *Justice, Political Liberalism, and Utilitarianism: Themes from Harsanyi and Rawls*, ed. Marc Fleurbaey, Maurice Salles and John A. Weymark (Cambridge, 2008), pp. 221–38, at 231, and that characterization is endorsed by Marc Fleurbaey, Maurice Salles and John A. Weymark, 'An Introduction to *Justice, Political Liberalism, and Utilitarianism*', *Justice, Political Liberalism, and Utilitarianism*, pp. 1–67, at 17. The line of objection originates in Amartya Sen, 'Welfare Inequalities and Rawlsian Axiomatics', *Theory and Decision* 7 (1976), pp. 243–62, and is developed and clarified in John A. Weymark, 'A Reconsideration of the Harsanyi–Sen Debate on Utilitarianism', *Interpersonal Comparisons of Well-Being*, ed. Jon Elster and John Roemer (Cambridge, 1991), pp. 255–316, and John Roemer, 'Harsanyi's Impartial Observer Is *Not* a Utilitarian', *Justice, Political Liberalism, and Utilitarianism*, pp. 129–35. There is also helpful discussion in Fleurbaey, Salles and Weymark, 'An Introduction', pp. 17–20 and 30–2.

[9] The interpersonal comparisons require not merely cardinality but co-cardinality. Nonetheless, without cardinality neither the intrapersonal nor the interpersonal comparisons are possible.

https://doi.org/10.1017/S0953820815000278 Published online by Cambridge University Press

This line of objection to Harsanyi, although widely regarded as successful, has nonetheless met some resistance.[10] One way to understand the thrust of the critique raised by Sen, Weymark and Roemer is that Harsanyi makes an illicit shift from the legitimate claim that the axioms of vNM utility theory entail that an individual's preferences can be represented by a vNM utility function to the stronger, and unfounded, claim that only a vNM utility function is an acceptable representation of those preferences.[11] John Broome acknowledges this criticism, yet argues that what he calls the 'expectational concept' of degrees of preference, according to which the vNM utility function is the appropriate representation, can be defended on the grounds that it is more natural than any of the alternatives.[12] He concedes that Sen, Weymark and Roemer are correct to point out a gap in Harsanyi's argument, and that expected utility theory on its own will not generate anything like the result Harsanyi is after, but he suggests that we can fill this gap by supplementing the argument with what he takes to be a plausible additional premise.

Broome's appeal to naturalness here is itself controversial, and perhaps rightly so.[13] But for the purposes of this article, I am not going to worry about whether the standard objection from Sen, Weymark and Roemer fully derails Harsanyi-style arguments or whether that line of objection can be overcome, as Broome suggests. Ultimately, that debate involves issues about preference measurement and interpersonal comparisons of preference and well-being. These are important and interesting issues, and deep problems lurk not just for Harsanyi's argument but for utilitarianism as a whole if they cannot be adequately addressed. Nonetheless, I contend that these issues are not at the heart of the debate between Rawls and Harsanyi. There is, admittedly, a disagreement between the two here. Rawls rejects the idea that

[10] I will focus on the response offered in Broome, 'Can There Be'. An alternative way of defending Harsanyi's argument, by adopting a substantive account of well-being independent of individuals' preferences, is suggested in Weymark, 'A Reconsideration', pp. 307–15. A version of this strategy, focused on idealized desires, is developed at greater length in Mathias Risse, 'Harsanyi's "Utilitarian Theorem" and Utilitarianism', *Nous* 36 (2002), pp. 550–77.

[11] Thanks to an anonymous referee for suggesting this way of understanding the objection to Harsanyi. This is also the way the objection is characterized in Broome, 'Can There Be', pp. 230–1 and in Fleurbaey et al., 'An Introduction', p. 17.

[12] Broome, 'Can There Be', pp. 230–2. Although Broome defends Harsanyi on this point, he finds other aspects of Harsanyi's argument problematic. Nonetheless, I take Broome to be pursuing a line of argument and analysis closely related to and inspired by Harsanyi, and one that has the potential to pose similar sorts of problems for Rawls. See Broome, *Weighing Goods*.

[13] Doubts about Broome's approach are raised by Fleurbaey, Salles and Weymark, 'An Introduction', and John A. Weymark, 'Measurement Theory and the Foundations of Utilitarianism', *Social Choice and Welfare*, 25 (2005), pp. 527–55.

https://doi.org/10.1017/S0953820815000278 Published online by Cambridge University Press

well-being should be understood in terms of preference satisfaction, and further rejects the use of any substantive account of well-being, partly because he wants his theory to be neutral between competing views of well-being (or between competing comprehensive conceptions of the good) and partly because he is concerned about issues of interpersonal comparison.[14] These considerations lead him to identify an objectively measurable list of primary goods as the things to be distributed according to his difference principle.[15] Given that the list of primary goods is designed to facilitate interpersonal comparisons, if we suppose that Rawls's only complaint against utilitarianism involves concerns about interpersonal comparisons, Rawls should have no objection to a utilitarian distribution of primary goods. But a utilitarianism of primary goods is just a slight modification of classical utilitarianism or preference utilitarianism, and a far cry from Rawls's own view. As a result, I take it that the heart of the debate between Rawls and utilitarians like Harsanyi must lie elsewhere than in issues about interpersonal comparisons. And Rawls himself concurs. After one of his discussions of the problems surrounding interpersonal comparisons, he writes: 'I believe the real objection to utilitarianism lies elsewhere.'[16]

My aim will therefore be to identify and evaluate a Rawlsian response to Harsanyi that is not focused on issues of interpersonal comparisons, and instead home in on what I take to be a more fundamental dispute between the two. In order to do this, I will proceed, for the sake of argument, as if Harsanyi's argument can withstand the standard objection. One benefit of this approach is that the ensuing analysis has the potential to reveal, and evaluate, a Rawlsian response to Hare's informal argument that original position reasoning leads to utilitarian conclusions, which is an argument Rawls needs to address even if Harsanyi's formal version of the argument cannot be salvaged from the critique posed by Sen, Weymark and Roemer.

II. A FIRST RAWLSIAN RESPONSE: NO PROBABILITY JUDGEMENTS

One line of response to Harsanyi that has been offered by Rawls and his defenders involves pointing out that Rawls formulates his preferred version of the original position in a way that prevents Harsanyi's argument from getting started. This is because Rawls imposes what has been called a 'thick' veil of ignorance that not only prevents people from knowing which roles they occupy in society, but prevents them

[14] Rawls, *Theory*, pp. 78, 150, 281–5.
[15] Rawls, *Justice as Fairness*, pp. 57–61 and 127, and *Theory*, p. 284.
[16] Rawls, *Theory*, p. 78.

https://doi.org/10.1017/S0953820815000278 Published online by Cambridge University Press

from even knowing the likelihood of occupying different roles. It might seem as though Harsanyi's equiprobability postulate would naturally kick in at this point with the insistence that deliberators in the original position take themselves to have an equal chance of occupying each role. The problem is that Harsanyi's postulate only makes sense if we are thinking of roles in a very fine-grained way. In Harsanyi's analysis, each individual in a society occupies a unique role, as is evident both from his explanation of the postulate above and from the way these 'roles' function in his subsequent proof. But Rawls thinks of roles as something more like positions in the society or in the social economy. So, for Rawls, the 'roles' relevant to a feudal society are things like 'king', 'aristocrat' and 'peasant'. Rawls, of course, would have parties in the original position think of roles in a less coarse-grained way than this, but the key point is that the roles need not be set up such that there is anything close to an equal number of people occupying each role in the society. If we apply Harsanyi's equiprobability postulate under this conception of 'roles' or 'social positions', we get the crazy result that one should evaluate the prospects of living in a feudal society under the assumption that one has an equal chance of being a king, an aristocrat or a peasant.

That is obviously not what Harsanyi had in mind. And the remedy may seem equally obvious. Harsanyi would say that if we are going to simplify by thinking of a society as consisting of kings, aristocrats and peasants, and if we are going to gloss over the differences between being this peasant or that peasant, perhaps because differences among the lives of peasants are insignificant in comparison to the difference between the life of a peasant and that of an aristocrat or a king, then we need to consider the proportion of kings, aristocrats and peasants in society. The equiprobability postulate appropriately applied to individual 'roles' dictates that we weigh these more coarse-grained social roles according to the number of people who occupy them.

At this point, however, a standard Rawlsian response is to point out that Rawls's thick veil of ignorance makes it impossible for deliberators in the original position to know how many people will occupy each social role, and so makes it impossible to apply Harsanyi's equiprobability postulate in a reasonable way. Given that the postulate is essential for Harsanyi's analysis, the Rawlsian conclusion is that Harsanyi's attempt to get from Rawls's original position model to utilitarianism does not even get off the ground.

I take this Rawlsian response to be compelling, and to constitute a decisive objection to any Harsanyi-style attempt to derive standard utilitarian conclusions from Rawls's own original position model. But I think it would be a mistake to suppose that this is the end of the story.

https://doi.org/10.1017/S0953820815000278 Published online by Cambridge University Press

Instead, it shows that there is a fundamental disagreement between Rawls and Harsanyi over the question of how to imagine or construct the original position. After all, it certainly seems possible to imagine an original position with only a 'thin' veil of ignorance, and thereby allow deliberators in that original position to have access to the sort of information required to make probabilistic judgements about the likelihood of occupying different roles in society. One might attempt to cast doubt on this possibility by arguing that even if we specify the principles of justice governing a feudal society in great detail, there will still be no precise answer to the question: 'What is the proportion of peasants, aristocrats and kings in a society governed by these principles?' In so far as that is correct, the uncertainty generates some complications for Harsanyi's formal proof. But his fundamental line of thought does not require great precision here. As long as deliberators in the original position can conclude something like 'under this sort of feudal system, roughly 90 per cent of the population will be peasants, roughly 10 per cent will be aristocrats, and one individual will be the king', they will be able to implement Harsanyi's equiprobability postulate and perform a rough calculation comparing the principles of justice that support that feudal system with other possible principles.

We have, then, two competing original position models. Both include veils of ignorance that screen out any self-identifying information. In other words, deliberators in both are asked to choose principles of justice to govern a society in which they shall live, without knowing which individuals within that society they will be, or which roles within that society they will occupy. In one model, this is all that the veil of ignorance does. In the other, the veil is thickened to prevent the parties in the original position from even knowing the relative frequency of different social roles, or the percentages of individuals that lead different types of lives in the society. Harsanyi's argument is intended to show that an original position model with a thin veil leads to the endorsement of standard utilitarian principles of justice. But deliberation within an original position model with a thick veil is not susceptible to his line of analysis, or so I have claimed.[17] This raises the question: how are we to choose which model to adopt?

In settling this question, I think it is instructive to look at what Rawls says about the construction of the original position and the purpose of the veil of ignorance. He writes:

[17] I have been focusing on Harsanyi's equiprobability argument, but the same conclusion follows for his axiomatic argument. Absent the information needed to make probabilistic judgements about the frequency of different social roles, neither argument can achieve its aims.

https://doi.org/10.1017/S0953820815000278 Published online by Cambridge University Press

> [I]f we are reasonable, it is one of our considered convictions that the fact that we occupy a particular social position, say, is not a good reason for us to accept, or to expect others to accept, a conception of justice that favors those in that position. . . . To model this and other similar convictions, we do not let the parties know the social position of the persons they represent. The same idea is extended to other features of persons by the veil of ignorance.[18]

The thing to notice here is that Rawls's reason for adopting a veil of ignorance is to prevent original position deliberation from being inappropriately impacted by the bargaining power of deliberators representing advantaged groups or individuals. This strikes me as a compelling reason for building a veil into the original position model. But an original position with only a thin veil seems to accomplish this goal.

Why, then, thicken the veil of ignorance? One possible reason is expressed well by James Lenman, who concedes Harsanyi's claim that allowing original position deliberators access to probabilistic information leads to standard utilitarian conclusions, but thinks Rawls is justified in preferring an original position model that stymies Harsanyi's line of analysis by screening off such information. Here is Lenman's analysis:

> [Rawls's] basic rationale for [depriving the contracting parties of any information about the probability of being a given individual (or more precisely, in Rawls's case, of occupying any given position in society)] was that the principles that Rawlsian contracting parties will choose cohere better, or so at least he urged, with basic moral intuitions about freedom and social justice than do utilitarian principles. This can sometimes look to beginning students of Rawls rather like cheating, fiddling around with the initial conditions to get the desired result, but that is to misunderstand his methodology.[19]

The methodology Lenman refers to here is that of reflective equilibrium, whereby we are to give significant weight to intuitive judgements at all levels of generality or abstraction. Specific judgements about particular actions or particular social policies are to be balanced against abstract judgements. In this case the abstract judgements concern how to construct an original position model that will accurately capture the demands of fairness or justice.

I am, myself, deeply ambivalent about reflective equilibrium in general. But even if we endorse reflective equilibrium wholeheartedly as an overall methodological approach, I think we should be highly suspicious of invoking it here, in the way Lenman does. The problem

[18] Rawls, *Justice as Fairness*, p. 18. Rawls says essentially the same thing in his discussion of the veil of ignorance in *Justice as Fairness*, pp. 82, 87.

[19] James Lenman, 'Review of Allan Gibbard, *Reconciling Our Aims: In Search of Bases for Ethics*', *Notre Dame Philosophical Reviews* (2009), <https://www.ndpr.nd.edu/news/24177-reconciling-our-aims-in-search-of-bases-for-ethics/> (2014).

https://doi.org/10.1017/S0953820815000278 Published online by Cambridge University Press

is that Rawls explicitly identifies his primary argumentative task as demonstrating the superiority of his preferred principles of justice over straightforward utilitarian principles and over restricted forms of utilitarianism. Moreover, Rawls at least appears to take the original position argument to be an important element of his overall analysis.[20] Yet Lenman is suggesting that Rawls concede that an original position model with a thin veil of ignorance leads to standard utilitarian conclusions. And when asked why we should instead endorse the principles of justice that follow from an original position with a thick veil, Rawls's answer is supposed to be that the thick veil is needed to avoid precisely those standard utilitarian conclusions, which are taken to be so obviously false or repugnant that we are entitled to make their avoidance a guiding principle in our construction of the original position. For those already convinced that utilitarianism leads to repugnant conclusions, and that this constitutes legitimate grounds for rejecting utilitarianism, this line of reasoning may seem palatable. But the original position analysis, conceived in this way, constitutes no more (nor less) of an argument against utilitarianism than is offered by simply pointing out that utilitarianism leads to counterintuitive conclusions. It is the counterintuitive nature of the conclusions that are said to follow from utilitarianism that does all of the work, and the analysis of deliberation within the original position is essentially no more than a charade, given that the original position itself is subject to reconception or redescription in order to generate the preferred outcome. If this were the right way to understand Rawls's analysis, I think it would indeed be fair to accuse him of 'fiddling around with the initial conditions to get the desired result'.

The point here is not that one cannot make a convincing argument against utilitarianism by pointing out the counterintuitive implications of the view.[21] Such arguments undeniably pose serious challenges to utilitarianism and require response.[22] Rather, the point is that Rawls

[20] In part II of *Justice as Fairness*, Rawls offers some general considerations in favour of his preferred principles of justice, but reserves his 'more formal and official argument' (Rawls, *Justice as Fairness*, p. 39) to part III, which is entirely structured in terms of what it would make sense to choose as parties in the original position. Given that, and the fact that *Justice as Fairness* is described by Rawls as a restatement and clarification of his view (Rawls, *Justice as Fairness*, pp. xv–xviii), it is hard to resist the conclusion that Rawls takes the original position argument to be a central and substantive element of his overall analysis.

[21] For a classic version of this sort of argument, see Williams, 'A Critique', which highlights both counterintuitive conclusions that are supposed to follow from utilitarianism and counterintuitive features supposedly displayed by utilitarian deliberation even when it leads to intuitively plausible conclusions.

[22] For an extended argument that utilitarianism avoids most of the counterintuitive conclusions typically associated with it, and, perhaps more importantly, that the intuitive

https://doi.org/10.1017/S0953820815000278 Published online by Cambridge University Press

evidently takes himself to be doing something different. The primary function, for him, of the original position analysis is to present or construct an argument showing that his preferred principles of justice are superior to utilitarian ones. But if there are two possible original position models, one of which leads to utilitarianism and the other does not, and if the only rationale for preferring the latter over the former is to avoid the former's utilitarian conclusions, then thinking in terms of an original position cannot play any significant role in developing an argument against utilitarianism.

Moreover, it seems to me that there are good reasons, internal to the idea of the original position, for including only a thin veil of ignorance. This is particularly clear if we follow Gibbard and think of the original position in terms of a 'You'd have agreed' retort.[23] Gibbard's idea is to imagine that someone raises an objection to an existing social practice, on the grounds that other ways of doing things would be better for that person. Gibbard's claim is that one way to effectively rebut such an objection is to say, 'Before you knew how you in particular would turn out to be affected, you would have agreed to the practice – and for your own advantage.'[24] In so far as the claim made in the rebuttal is accurate, and so the original objector really would have agreed, Gibbard views this sort of retort as decisive. The point of an original position model is therefore to capture the moral force of this 'You'd have agreed' retort.

As Gibbard recognizes, this is just another way of expressing the same basic idea that Rawls identifies as a motive for engaging in original position analysis. But unlike Rawls, Gibbard adopts a Harsanyi-style original position with a thin veil of ignorance. Moreover, although he does not address this issue directly, it strikes me that Gibbard's way of thinking about the motivation for engaging in original position reasoning reveals a strong basis for adopting an original position model that allows deliberators access to the information necessary for probabilistic reasoning.

To see this, consider the following toy case. Imagine a situation with ten people and three outcomes. Call the outcomes A, B and C. Suppose all ten people agree that A is better than B, which is better than C. Suppose there is also a consensus that having a 90 per cent chance of A and a 10 per cent chance of C is better than a 100 per cent chance of B. Moreover, there are two competing policies the group could

plausibility of conclusions of this sort is not dispositive of utilitarianism, see Hare, *Moral Thinking*.

[23] Allan Gibbard, *Reconciling Our Aims: In Search of Bases for Ethics* (Oxford, 2008), pp. 50–1 and 150–2.

[24] Gibbard, *Reconciling*, p. 50.

https://doi.org/10.1017/S0953820815000278 Published online by Cambridge University Press

adopt: X and Y. If they adopt X, nine will get A and one will get C. If they adopt Y, all will get B. Now suppose that one of these people is placed in a Harsanyi-style original position with a thin veil of ignorance. Reasoning as Harsanyi suggests, this person chooses policy X, because the 90 per cent chance of A and 10 per cent chance of C it provides is better than the 100 per cent chance of B provided by policy Y.

The veil is then lifted, and the chooser is lucky enough to be one of the nine who gets A. The one who ended up with C seeks out the chooser and makes the following complaint: 'Why did you choose X? If you had chosen Y I'd have got B rather than C. It's not fair! You and the others have benefited at my expense! You should have chosen Y instead.' This may seem like a compelling objection, but notice that the chooser can respond, 'But if you had been choosing, not knowing who you were, you would have chosen X as well', thereby invoking the 'You'd have agreed' retort.

Given the stipulations built into the example, the one raising the objection cannot legitimately deny what is asserted in this retort. But imagine that the objector tries to undermine the moral force of the retort by saying, 'Certainly I would have agreed to X had I been in your choice situation, but that doesn't matter. I only would have chosen X on the basis of probabilistic reasoning. What matters is whether I would have chosen X or Y if I didn't have the information to reason probabilistically, and only knew that Y would give everyone B, whereas X would give each of us either A or C. In that case I'd have chosen Y. That's why X is unfair.' Call this the 'Only on the basis of probabilistic reasoning' counter-retort.

This is, to some extent, a stand-off. If one side says 'You'd have agreed' and the other says, 'Only on the basis of probabilistic reasoning', I am not sure what else either side could add to strengthen its position. By my lights, though, this is a stand-off with a clear winner. In other words, I take the 'You'd have agreed' retort to have an almost undeniable moral force, although I share Gibbard's professed inability to offer a definitive argument for that view. As Gibbard suggests, perhaps the best that can be done is to state as clearly as possible what would be involved in denying it.[25] The idea would be that someone, call him Jack, objects to a principle of justice on the grounds that it does not adequately

[25] Gibbard, *Reconciling*, p. 152, writes: 'if the retort leaves cold someone who genuinely understands what it involves, then I don't know anything to say that would make the person responsive.' As Gibbard, *Reconciling*, p. 50 n. 30, notes, Broome and Brian Barry report being unmoved by this general line of thought in Broome, *Weighing Goods*, pp. 55–9 and Brian Barry, *Theories of Justice* (Berkeley, 1989), pp. 334–5. Perhaps, though, it is significant that they are responding to the idea as formulated by Harsanyi rather than to Gibbard's own formulation in terms of a 'You'd have agreed' retort, which seems to me to highlight the moral force of the idea.

https://doi.org/10.1017/S0953820815000278 Published online by Cambridge University Press

take his interests into account. And then Jack concedes that he would endorse the principle if he were evaluating it with those very same sorts of interests in mind, but not knowing who he was, or, more precisely, believing himself to have an equal chance of being anyone. In spite of this concession, Jack maintains his objection, grounded the same way as before, in the claim that his interests were not adequately taken into account. Is this not a clear case of Jack being unreasonable? I find it extremely difficult to view it as anything other than that, which is to say that I find the 'You'd have agreed' retort to have genuine moral force, although, as indicated above, I am unsure of what more to say to those who see things otherwise.

In contrast, the 'Only on the basis of probabilistic reasoning' counter-retort leaves me mostly bewildered, and I am puzzled about how to understand or engage with someone who takes it to be an appropriate counter to the 'You'd have agreed' retort. One possible explanation, I suppose, for why one might take this counter-retort to be compelling might be that one has a pre-existing commitment to rejecting standard utilitarian principles of justice, for independent reasons. If those reasons are sufficiently strong, and if one believes that taking 'Only on the basis of probabilistic reasoning' as a compelling response to 'You'd have agreed' is necessary to avoid utilitarian conclusions, then one might be able to rationalize viewing the counter-retort favourably. But this line of thought is just another way of adopting Lenman's invocation of reflective equilibrium, which I have argued robs Rawls's original position analysis of any genuine force against utilitarianism.[26] If the original position argument is to play the role Rawls assigns to it, then this cannot be an appropriate basis for accepting the 'Only on the basis of probabilistic reasoning' counter-retort.

Another possible way of defending the 'Only on the basis of probabilistic reasoning' counter-retort is to appeal to Rawls's analysis from *A Theory of Justice*, which contains a line of argument that does not appear in *Justice as Fairness*. There Rawls focuses on the principle of insufficient reason, which tells us that if we have no information on which to base probability judgements we should assign an equal probability to each possible outcome. Rawls rejects the use of the principle of insufficient reason in the original position because of 'the fundamental importance of the original agreement and the desire to have one's decision appear responsible to one's descendants who will be

[26] As indicated above, this is not to say that invoking reflective equilibrium in this way makes it impossible to argue against utilitarianism, just that it makes it impossible for that argument to be grounded, in any meaningful way, in an original position analysis.

https://doi.org/10.1017/S0953820815000278 Published online by Cambridge University Press

affected by it'.[27] His concern is that probability judgements grounded solely in the principle of insufficient reason are less reliable than those grounded in information about the world. As a result, given the high stakes of original position reasoning, it is inappropriate or irresponsible to allow them to play a significant role in one's deliberation.[28]

As I see it, the problem with this line of thought is that it mischaracterizes what is at issue. The question is not whether original position deliberators who lack any basis for making probability estimates should rely on the principle of insufficient reason. Rather, the question is how we should characterize the original position, or what version of the original position we should care about. If Rawls's analysis is compelling, all it shows is that when we imagine deliberators in an original position designed in such a way that they have no basis for knowing whose role in society they will occupy, we should not allow them to assume they will have an equal chance of being in each role. But we can understand an original position with a thin veil of ignorance as one in which it is stipulated that the deliberators have an equal chance of being anyone. In that sort of original position, the principle of insufficient reason need not be invoked, and Rawls's argument against relying on it is irrelevant.

Think again of the toy example discussed earlier. As it turns out, nine people get A and one gets C. Had their society been designed otherwise, all would have got B. The one who gets C complains about the institutional design of their society. We want to know whether the complaint is reasonable. The argument for it being unreasonable is grounded in the following uncontested assertion: (1) had that person, or any of the others, been in a position to choose the societal design, knowing that they had an equal chance of being any of the ten, they would have chosen the existing design. The argument for it being reasonable is grounded in a separate, and also uncontested, assertion: (2) had that person, or any of the others, been in a position to choose the societal design, not having any basis for estimating the likelihood of getting A or C under the existing design, they would have chosen the other design.[29] I am willing to concede that (2) provides some support for the conclusion that the complaint is reasonable. In the absence of (1), or anything like it, (2) might even be enough to settle the issue. But when (1) and (2) are both acknowledged, it seems evident to me

[27] Rawls, *Theory*, p. 146. Thanks to an anonymous referee for calling my attention to this quote, and the related line of argument.

[28] Rawls, *Theory*, pp. 149–50.

[29] I am stipulating that these assertions are uncontested in order to focus attention on a clear case in which opting for an original position with a thin veil of ignorance leads to a different conclusion from opting for one with a thick veil of ignorance.

https://doi.org/10.1017/S0953820815000278 Published online by Cambridge University Press

that the conclusion grounded in (1), according to which the complaint is unreasonable, supersedes the conclusion grounded in (2). Admittedly, this is just another way of asserting my view that the 'You'd have agreed' retort is morally compelling and the 'Only on the basis of probabilistic reasoning' counter-retort is not. I still have no definitive argument for that view, although I do think I have shown why Rawls's discussion of the principle of insufficient reason cannot serve as the basis of an argument against it. The fundamental problem is that Rawls's analysis focuses on the epistemic propriety of adopting the principle of insufficient reason, whereas what is actually at issue is how to characterize the original position in order to ensure that choices made within it are appropriately fair.

III. A SECOND RAWLSIAN RESPONSE: AVOID INTOLERABLE OUTCOMES

Perhaps in part for the reasons described above, Rawls himself often seems reluctant to rely on the idea that parties in the original position have no access to the information required for probabilistic reasoning.[30] Instead, he focuses primarily on three other conditions that he says obtain in the original position and appropriately guide the reasoning of deliberators in that position.[31] These conditions all involve the outcomes associated with the competing sets of principles of justice between which parties in the original position are choosing. The first condition is that the parties can identify a set of principles of justice such that the worst possible outcome under those principles is acceptable, or satisfactory. Rawls calls this outcome the 'guaranteeable level', because if those principles are chosen then everyone is guaranteed an outcome at least that good. The second condition is that any outcomes superior to this guaranteeable level are only marginally superior. The important thing about this second condition is that it applies to any outcomes under any set of principles, not merely the outcomes that are possible under the principles that generate the satisfactory guaranteeable level. The third condition is that under any alternative set of principles there is at least one outcome that Rawls characterizes as 'intolerable'.

[30] See, for instance, Rawls, *Justice as Fairness*, p. 101 (in the context of the first fundamental comparison) and p. 120 (in the context of the second).

[31] These conditions are formulated on Rawls, *Justice as Fairness*, p. 98. My presentation of the list is different from his in two ways, but the substance is the same. The differences are as follows. (1) He lists the lack of information required for probabilistic reasoning as the first of three conditions. That does not appear on my list because I am, at this point, attempting to capture only his other conditions. (2) He combines the first and second conditions on my list. Together, they constitute the second condition on his list. I have separated them in an effort to ensure clarity in my subsequent analysis.

https://doi.org/10.1017/S0953820815000278 Published online by Cambridge University Press

Rawls thinks of these conditions as capable of obtaining to varying degrees. When they obtain fully, as he claims happens when we take his preferred principles of justice and straightforward utilitarian principles as the only two options under consideration, Rawls argues that it makes sense to choose the principles that generate the guaranteeable level. Rawls famously presents this argument as operating through the maximin rule of choice, which instructs one to identify the worst possible outcome under each set of principles of justice being considered and then choose whichever set generates the most desirable worst outcome. Constructed in terms of this maximin rule, Rawls's argument has three steps. First, he argues that the conditions mentioned above obtain. Second, he argues that when they obtain it is appropriate to adopt a maximin rule of choice. Third, he applies the maximin rule to select his preferred principles over the utilitarian alternative.

In my view, though, it is easier to understand Rawls's core argument if we simplify the analysis and avoid mention of a generalized maximin choice procedure.[32] The fundamental idea is that we should choose principles of justice that rule out intolerable outcomes over those that allow such outcomes, especially if any potential gains provided by the principles permitting intolerable outcomes are minimal. When the conditions mentioned above obtain, there is only one available option that rules out the possibility of intolerable outcomes, because the first condition ensures that there is such an option and the third ensures that it is the only such option. When that occurs, and perhaps especially when the second condition further ensures that any potential gains provided by principles that permit intolerable outcomes are minimal, the idea that it makes sense to prefer principles that rule out intolerable outcomes to those that allow them is sufficient to determine the choice made in the original position.

But what exactly grounds this decisive preference for avoiding the possibility of intolerable outcomes? In considering this question, it is

[32] This also avoids creating the false impression that Rawls wants to derive the difference principle from the maximin rule of choice. Rawls, *Justice as Fairness*, p. 43 n. 3 talks about this misconception, and admits his own role in creating it. My preferred approach also avoids creating the impression that Rawls is advocating the general use of a maximin principle in deliberation, which he is not, as indicated in Rawls, *Justice as Fairness*, p. 97 n. 9 and p. 99. It is also worth noting that Rawls, *Justice as Fairness*, p. 99 acknowledges that the maximin principle is not essential to his analysis, and that a presentation like the one I am offering fully captures his line of thought. Given the extent to which framing the argument in terms of the maximin principle has led to misunderstanding and confusion, and the fact that the argument can be presented more simply without appeal to the maximin principle, I think it is preferable to work with a version of Rawls's view that avoids invoking maximin reasoning.

https://doi.org/10.1017/S0953820815000278 Published online by Cambridge University Press

instructive to notice what Harsanyi would say here. On his approach, the so-called 'intolerable' outcomes can be assigned utility values, just like any other outcome. Calling them intolerable is therefore just a way of emphasizing that they get assigned exceptionally low values. But no matter how low they are, Harsanyi will insist that it does not make sense for parties in the original position to adopt a strict policy of preferring any principles that rule them out over any principles that allow them. Rather, it will depend on the other outcomes. If we restrict our attention to cases in which the second condition obtains fully, which is to say cases in which the principles allowing intolerable outcomes generate only small potential benefits relative to the outcomes realized for everyone under the principles that avoid anything intolerable, we are closer to having a conclusive argument in favour of principles that avoid intolerable outcomes. But even for these cases, according to Harsanyi, we are not warranted in adopting a strict policy of preferring principles that rule out intolerable outcomes. As he sees it, if the chance of ending up with an intolerable outcome is sufficiently small, then a sufficiently large likelihood of getting a minimally better outcome can rationalize the choice of principles that allow the intolerable. This, he thinks, follows from the fairly weak Bayesian rationality conditions needed to assign utilities to outcomes associated with roles or individuals and then deliberate under uncertainty about which role one will occupy, or which individual one will be.

One way to block arguments like Harsanyi's, of course, is to deny access to the information necessary for probabilistic reasoning. But if my analysis in the previous section is correct, that way of responding to Harsanyi should be ruled out, because there are good reasons for endorsing the principles chosen from an original position model that allows the possibility of probabilistic reasoning.[33] There is, though, another way of responding to Harsanyi, one which I think Rawls and advocates of a Rawlsian analysis have overlooked, and which I think has some potential to vindicate elements of Rawlsian principles of justice over standard utilitarian ones. What I have in mind is the possibility of objecting to Harsanyi's way of cashing out the idea of 'intolerable' outcomes, which will involve calling into question one of the supposedly weak Bayesian rationality conditions on which Harsanyi's analysis depends.

[33] As discussed in sect. I, one could also block Harsanyi's argument by appealing to the standard objection posed by Sen, Weymark and Roemer. That may very well be an effective response to Harsanyi, but, as indicated earlier, it does not shed light on what I take to be the core disagreement between Rawls and Harsanyi.

https://doi.org/10.1017/S0953820815000278 Published online by Cambridge University Press

IV. HARSANYI'S BAYESIAN RATIONALITY CONDITIONS

As indicated earlier, the key to Harsanyi's analysis is the construction of von Neumann–Morgenstern utility functions, or vNM utility functions, which represent the value an agent assigns to possible outcomes by looking at the agent's preferences between lotteries that offer different probabilities of ending up with different mixtures of these outcomes. Harsanyi acknowledges that this approach depends on several Bayesian rationality postulates, and, in 'Morality and the Theory of Rational Behavior', he defends that reliance as follows:

> Most philosophers and social scientists do not realize how weak the rationality postulates are that Bayesian decision theory needs for establishing the expected-utility maximization theorem. . . . [A]ll we need is the requirement of consistent preferences (complete preordering), a continuity axiom, the sure-thing principle (avoidance of dominated strategies), and the requirement that our preferences for lotteries should depend only on the possible prizes and on the specific random events deciding the actual prize. (The last requirement can be replaced by appropriate axioms specifying the behavior of numerical probabilities within lotteries. In the literature, these axioms are usually called 'notational conventions'.)[34]

The sure-thing principle, which is closely related to the principle of independence, has been subject to significant debate, primarily in connection with the Allais paradox and the Ellsberg paradox. One avenue an advocate of Rawls might take is therefore to challenge Harsanyi's sure-thing principle. I will not explore that here, partly because I think others have offered convincing defences of the sure-thing principle.[35] More importantly, though, I will focus elsewhere because I think there is another avenue of potential objection on Rawls's behalf, and, unlike a possible rejection of the sure-thing principle, it is one that strikes me as capturing the heart of Rawls's line of analysis. From Rawls's perspective, I think the truly problematic item in Harsanyi's list of assumptions is the continuity axiom.

In 'Morality and the Theory of Rational Behavior', Harsanyi does not give a precise account of the continuity axiom, but he refers back to an earlier article in which he defines the continuity axiom necessary for his analysis as follows:

> Postulate II (continuity): If prospect P is preferred to prospect R, while prospect Q has an intermediate position between them (being preferred to R but less

[34] Harsanyi, 'Morality', p. 636 n. 17.

[35] For a discussion of the sure thing principle in connection with the Allais paradox and related objections, see Broome, *Weighing Goods*, pp. 94–117. For a discussion of the sure thing principle in connection with the Ellsberg paradox, see Ken Binmore and Alex Voorhoeve, 'Defending Transitivity against Zeno's Paradox', *Philosophy and Public Affairs* 31 (2003), pp. 272–9.

https://doi.org/10.1017/S0953820815000278 Published online by Cambridge University Press

preferred than P), then there exists a mixture of P and R, with appropriate probabilities, such as to be exactly indifferent to Q.[36]

It is perhaps worth noting that in this earlier presentation of the Bayesian rationality postulates Harsanyi calls this continuity axiom 'almost trivial', even though he acknowledges some room for possible debate over the postulate he uses in the place of the sure-thing principle. Similarly, in his extensive analysis of Harsanyi's arguments, Broome classifies continuity as a 'technical axiom' and says that he is 'simply going to take [it] for granted'.[37] On closer inspection, though, I think the continuity axiom turns out to be far from trivial, or something to be taken for granted, and in fact is at the heart of a fundamental disagreement between Rawls and utilitarians like Harsanyi.

What exactly does the continuity axiom say? Using the variables from Harsanyi's definition, the idea is that we are to imagine that one is faced with three possible outcomes, P, Q and R, and that one judges P to be better than Q and Q to be better than R. In other words, if faced with a choice between a 100 per cent chance of P and a 100 per cent chance of Q, one chooses a 100 per cent chance of P, and if faced with a choice between a 100 per cent chance of R and a 100 per cent chance of Q, one chooses a 100 per cent chance of Q. What, though, if one is faced with the following choice?

(1) 100 per cent chance of Q
or (2) 50 per cent chance of P and 50 per cent chance of R

In this case, there is not yet enough information to determine one's choice. It will depend on the relative values one assigns to P, Q and R. Or more precisely, it will depend on comparing the difference in the values one assigns to P and Q with the difference in the values one assigns to Q and R. Suppose one thinks P is sufficiently appealing relative to Q, and R is sufficiently acceptable relative to Q, that one chooses (2) over (1). What the continuity axiom asserts is that as we modify (2) by decreasing the likelihood of P and increasing the likelihood of R, we will eventually come to a tipping point. There will be some probability, call it x, such that one is indifferent with respect to the following choice:

(1) 100 per cent chance of Q
(3) probability x of P and probability $(1 - x)$ of R

This x may be very low, but according to the continuity axiom there will always be some x such that one is indifferent between (1) and (3).

[36] Harsanyi, 'Cardinal Welfare', p. 313 n. 10.

[37] Broome, *Weighing Goods*, p. 94. Similarly, Risse takes continuity to be potentially controversial only in the context of comparing the costs and benefits of eternal damnation or salvation with earthly goods. See Risse, 'Harsanyi's "Utilitarian Theorem"', p. 561.

https://doi.org/10.1017/S0953820815000278 Published online by Cambridge University Press

Similarly, if one thinks Q is sufficiently good relative to P and R that one chooses (1) over (2), it will be possible to modify (2) by increasing the likelihood of P and decreasing the likelihood of R, eventually reaching a point of indifference.

The relevant upshot of this for the debate between Rawls and Harsanyi is that the continuity axiom has the following implication. No matter how bad one judges R to be, and no matter how slight one's preference is for P over Q, one will always be willing to risk the possibility of R (rather than Q) in order to secure the benefits of P (rather than Q), provided that the risk of ending up with R is sufficiently small.[38]

V. INTOLERABLE OUTCOMES AND CONTINUITY

It should be clear that the continuity axiom is a central element in the analysis of 'intolerable' outcomes that I earlier offered on Harsanyi's behalf. The continuity axiom is what ensures that parties in the original position can assign even an intolerable outcome a score on a continuous utility scale, and it is the continuity of that utility scale that prevents them from strictly ruling out principles of justice that permit such an intolerable outcome in favour of those that ensure the avoidance of any outcomes that are intolerable.

As I see it, the proper Rawlsian response here is to claim that this misconstrues what is involved in characterizing an outcome as intolerable. Rather than interpreting 'intolerable' as meaning something like 'very bad', the Rawlsian ought to say that identifying an outcome as intolerable involves positing a discontinuity in the scale of evaluation, thereby rejecting the continuity axiom. On this suggestion, to say that an outcome is intolerable is to commit oneself not to accept even a minimal risk of that outcome in order to achieve a gain within the realm of the tolerable. In terms of the 'You'd have agreed' retort, the claim here is that there are some things so bad that one would not have agreed to any chance at them, no matter how small, in order to achieve a gain within the realm of the tolerable.

[38] Strictly speaking, the continuity axiom on its own is only sufficient to ensure that there is some very small risk of R such that I am indifferent between running that risk or not running it. The conclusion as stated in the text also requires use of the sure thing principle. Stating it as I have done highlights the potential Rawlsian objection, and it should be clear, even to those who are suspicious of the sure thing principle in general, that the continuity axiom is what is at issue here. (The only work being done here by the sure thing principle is to license the idea that if P is preferable to R, and one is comparing lotteries with P and R as the only outcomes, a lottery with a higher probability of P and a lower corresponding probability of R is preferable to one with a lower probability of P and a higher corresponding probability of R.)

https://doi.org/10.1017/S0953820815000278 Published online by Cambridge University Press

Under this understanding of intolerability, it really does make sense for parties in the original position to have a strict preference for all principles that rule out intolerable outcomes over any that permit them. If the first and third of Rawls's conditions mentioned above are satisfied, so that there is a tolerable guaranteeable level, and all sets of principles other than the one which provides this guarantee contain the possibility of intolerable outcomes, that is sufficient to ground an original position preference for the tolerability-ensuring set of principles.[39] Harsanyi's line of analysis has no force here, because the Rawlsian view I am outlining involves the rejection of one of Harsanyi's key premises.

Let me be clear that I do not mean to suggest that Rawls conceives of himself as positing discontinuities of value and rejecting Harsanyi's continuity axiom. Rather, my claim is that this is a way for Rawls to maintain his commitment to an original position analysis and nonetheless avoid the conclusions Harsanyi attempts to force upon him, regardless of whether Harsanyi's argument can be saved from the standard objection discussed earlier, and regardless of whether utilitarians can solve the problems of interpersonal utility comparisons. I also contend that this way of blocking Harsanyi is very much in the spirit of Rawls's overall line of analysis, especially his emphasis on avoiding intolerable outcomes.

VI. DISCONTINUITIES OF VALUE

On the line of analysis I am suggesting, everything hinges on whether there really are discontinuities of value that make it possible for outcomes to qualify as intolerable. In thinking about this, it is worth noticing that intolerability, as I am proposing we understand it, is a relative notion. Y is intolerable relative to X if there is a discontinuity of value between them. But if Z is also below the point of discontinuity, then Y is not intolerable relative to Z, and it is quite possible for Y to be better than Z, perhaps even much better, even if both are intolerable relative to X. Moreover, if there are multiple points of discontinuity, it is possible that Y and Z are intolerable relative to X, and all three are intolerable relative to W.

[39] Rawls' second condition, which is that outcomes above the guaranteeable level are only marginally above it, ends up being unnecessary. As I see it, this reveals some equivocation on Rawls's part between two lines of argument. On the one hand, there is an argument like the one I am developing here on his behalf, in which discontinuity is the key to rejecting a standard utilitarian analysis. In that argument, Rawls's second condition is superfluous. On the other hand, there is an argument that I would characterize as an attempt to ground Rawls's preferred principles in standard utilitarian considerations (although Rawls would not like that characterization), and in that argument the second condition is relevant and important.

https://doi.org/10.1017/S0953820815000278 Published online by Cambridge University Press

From within an original position model, even one with a thin veil of ignorance that permits probabilistic reasoning, these points of discontinuity allow the attainment of outcomes above a point of discontinuity to trump other considerations. If we keep things relatively simple and imagine a single point of discontinuity, and if we continue to suppose that a line of analysis like Harsanyi's can survive the standard objection, the original position decision procedure would be as follows. First, check if there are any options that have any chance of an outcome above the discontinuity. If not, apply a standard expected utility calculation, trading off the prospects of better or worse outcomes below the point of discontinuity against one another. But if there are options that have any chance of an outcome above the discontinuity, then simply evaluate each option based on the chance it provides of such outcomes. All that matters is increasing the odds of being above the discontinuity.[40] And if there are any options that guarantee being above the discontinuity, rule out all others and evaluate these options against one another through a standard expected utility calculation.

It is important to recognize that the existence of discontinuities in value would make this style of deliberation appropriate not merely within the original position but in real-world self-interested deliberation as well, or at least in any such deliberation involving the possibility of intolerable outcomes, provided the agent were to recognize the discontinuity and were reasoning well. One way to evaluate the plausibility of value discontinuity is therefore to consider the implications it would have for ordinary self-interested deliberation.

The first thing to notice here is that this is not the way in which deliberation typically proceeds. As Harsanyi and others have pointed out, we routinely take on significant risks, even the risk of death, and often when the potential gain is relatively low. Moreover, not only do we do this, for instance when we cross the road to reach a restaurant that is slightly preferable to the one beside us, but it seems perfectly reasonable for us to do so, provided the risk is sufficiently small.[41] Our everyday behaviour therefore commits us to the view that death is not discontinuously worse than life.[42] And if death is not

[40] If the chance of being above the discontinuity were tied, then the relative evaluation of outcomes within the tiers would become significant.

[41] The example of crossing a road is one of several of this sort mentioned in John C. Harsanyi, 'John Rawls' Theory of Justice: Some Critical Comments', *Justice, Political Liberalism, and Utilitarianism*, pp. 71–9, at 72. A similar example involving plane travel is discussed at greater length in Harsanyi, 'Can the Maximin Principle', p. 595. See also Alastair Norcross, 'Great Harms from Small Benefits Grow: How Death Can Be Outweighed by Headaches', *Analysis* 58 (1998), pp. 152–8.

[42] To avoid verbiage, if there is a point of value discontinuity and X is below it and Y above it, I will call X 'discontinuously worse' than Y.

https://doi.org/10.1017/S0953820815000278 Published online by Cambridge University Press

discontinuously worse than life, it may seem as though the prospects for value discontinuity are dim indeed.

Nonetheless, I think there are recognizable situations, or at least literary tropes, in which the style of deliberation called for by judgements of discontinuity seems to occur. Consider, for instance, the love-struck paramour who risks everything for the chance to be with his or her beloved. Or imagine the adrenaline junky who risks everything for one more thrilling experience. These look like cases of people who are trying to achieve something judged to be discontinuously better than the alternative. That discontinuity is what potentially rationalizes their behaviour. But notice the implications of positing a true discontinuity here. If life with one's beloved is discontinuously better than life without, it is rational to risk everything else one cares about for even the slightest chance of that life. Thinking that way may have romantic appeal, but keep in mind that this involves much more than asserting that one cares more about being with one's beloved than about anything else, or even everything else put together. Those sorts of claims are perfectly compatible with value continuity. Moreover, value continuity is also perfectly compatible with the idea that it may make sense to give up everything else one cares about to be with one's beloved, or even just to have a chance to be with one's beloved. What value continuity requires, though, is that the size of that chance matters. Value continuity requires that there be some point, as that chance becomes increasingly slim, where the risk no longer makes sense. What was a noble act of love eventually gets transformed into a blind act of futility. Discontinuity, on the other hand, requires that the risk make sense even if one is making a substantial sacrifice in order to gain only a marginal chance of success. On this view, love is noble to the bitter end. There is a romantic, and I think understandably appealing, conception of love that involves this sort of discontinuity of value. And it may even be that it makes sense to be the kind of person who acts as if there is a discontinuity of value here.[43] But I find it difficult to maintain the view that large genuine sacrifices make sense for marginal increases in the likelihood of being with one's beloved. At some point the noble and admirable act of love really does become a stubborn and blind act of futility, or so it seems to me.

This line of thought does not rule out the possibility of discontinuities in value, but it does, I think, suggest that they are uncommon, at best.[44]

[43] The possibility that it could make sense to act as if there is a value discontinuity even if we suppose that there is not actually any such discontinuity is suggested by the lines of thought developed in Hare, *Moral Thinking*, and in Railton 'Alienation'.

[44] In contrast, Dale Dorsey has recently argued that discontinuity is a relatively common phenomenon. See Dale Dorsey, 'Headaches, Lives and Value', *Utilitas* 21 (2012),

https://doi.org/10.1017/S0953820815000278 Published online by Cambridge University Press

Or more precisely, it emphasizes the radical nature of positing value discontinuities. Were value discontinuity a common phenomenon, it would require radical revision to everyday deliberative practices.

VII. DISCONTINUITY AND RAWLS'S PRINCIPLES OF JUSTICE

A Rawlsian, though, does not need to assert that Harsanyi's continuity axiom is routinely violated. A relatively small set of discontinuities would be sufficient to lead parties in the original position to endorse Rawlsian principles of justice. If I am correct that these potential discontinuities are the key to a fundamental disagreement between Rawls and Harsanyi, it is important to consider where they would lie.

As indicated earlier, Rawls's preferred principles of justice require (1) equal basic liberties for all, (2a) fair equality of opportunity, and (2b) that inequalities be arranged to be to the benefit of the worst-off, as demanded by the difference principle. Rawls further takes each of these requirements to be lexically prior to the next, which is to say that his principles can be understood as positing three points of discontinuity. The first is between having one's equal basic liberties and not having them. The second is between having one's equal basic liberties with fair equality of opportunity and without it. The third is between having one's equal basic liberties and fair equality of opportunity with inequalities governed by the difference principle and having equal basic liberties and fair equality of opportunity without the difference principle.

I think there is good reason for at least some scepticism about all of these posited discontinuities, but let me initially focus on the third, which I think is the most problematic. As I see it, Rawls's own analysis in support of the difference principle highlights the difficulty he has here. He frames his discussion as a choice between his preferred principles and what he calls the principle of restricted

pp. 36–58. In particular, he claims that there is a discontinuity between the value of fulfilling what he calls a 'deliberative project', on the one hand, and satisfying 'momentary desires' or realizing 'hedonic achievements', on the other. His primary examples of deliberative projects are becoming a philosopher or a trombone player, and his primary example of momentary desires and hedonic achievements involves avoiding a mild headache. Dorsey's argument is complicated, and his conception of continuity is not exactly the same as mine, which further complicates matters. His analysis therefore requires a more detailed treatment than I can provide here. Briefly, though, my view is that the analysis from the text applies to his preferred examples as well. If I am told that the only way of relieving my headache poses a risk of interfering with one of my deliberative projects, that will indeed deter me from relieving the headache. But as the risk becomes increasingly slim, I think it eventually makes sense to run the risk for the sake of alleviating the headache. I take it that anyone who takes recourse to ibuprofen or acetaminophen to treat a headache concurs, or at least acts as if they concur.

https://doi.org/10.1017/S0953820815000278 Published online by Cambridge University Press

utility, which replaces the difference principle with a guaranteed social minimum and a utilitarian principle of distribution beyond that. He quickly concedes that both sets of principles provide a tolerable guaranteeable level, which is to say that both ensure the avoidance of any intolerable outcomes.[45] Given that I have been interpreting Rawls's claims of intolerability as proposed violations of Harsanyi's continuity axiom, it might therefore seem to be a mistake to even construe Rawls's argument for the difference principle in terms of discontinuity. Nonetheless, I think Rawls's ensuing line of analysis is naturally interpreted as aiming for something like a discontinuity claim. And interpreting Rawls in this way at least allows for the possibility that he could effectively block arguments like Harsanyi's.

Rawls proceeds to formulate his defence of the difference principle in terms of reciprocity and stability.[46] The central idea is to introduce a reasonable agreement test, under which the parties in the original position ask themselves whether the members of a society governed by a particular set of principles of justice would be able to endorse those principles and justify them to one another. The ability to do this, Rawls claims, is limited by what he calls the strains of commitment. The idea is that in any society there will be some citizens who would be better off if the social structure were governed by different principles of justice. For the society to be appropriately stable, these people must be able to remain committed in spite of this strain. Rawls then claims that in a society governed by any set of principles of justice that do not incorporate the difference principle, or something very similar to it, the least advantaged will suffer from excessive strains of commitment that will prevent them from endorsing the principles in question. Imagining himself as one of these people, he writes: 'We feel left out; and, withdrawn and cynical, we cannot affirm the principles of justice in our thought and conduct over a complete life.'[47] The reason for this, according to Rawls, is that it will seem to the least advantaged as if their interests have not been adequately taken into account, given that the conditions of strict reciprocity embedded in the difference principle have not been realized.[48] Expressed as a discontinuity claim, the idea here is that the strains of commitment suffered by the least advantaged under principles of justice that do not include the difference principle

[45] Rawls, *Justice as Fairness*, p. 120.
[46] Rawls, *Justice as Fairness*, pp. 122–30.
[47] Rawls, *Justice as Fairness*, p. 128.
[48] Rawls's analysis depends at this point on an empirical, psychological claim, and scepticism about this claim may very well be warranted. But for the sake of argument I am going to grant Rawls his psychological claim and attempt to show that his argument does not succeed even if he is correct about this.

https://doi.org/10.1017/S0953820815000278 Published online by Cambridge University Press

are so severe that it is not worth risking any chance at them in order to achieve mere economic gains.

The problem with this line of analysis is that it fails to acknowledge the straightforward way in which the interests of the least advantaged are taken into account under the principle of restricted utility. To see this, it may help to think again in terms of the 'You'd have agreed' retort. According to Rawls, the legitimate complaint registered by the least advantaged, provided the difference principle is not being followed, is something like, 'I would be better off if we adhered to the difference principle, and the decision not to do so seems not to have taken my interests into account properly, and that is unacceptable.' In response, others can say, 'But if you hadn't known who you were, and had those sorts of interests in mind, you'd have agreed to the current system.' At this point, the only real avenue of response is for the least advantaged to say, 'No I wouldn't have. I would have known it would seem to me that my interests weren't adequately taken into account, and having things seem that way is so bad that I'd never have agreed to any risk of that.'

This sort of response from the least advantaged, if legitimate, would indeed be effective. It would show that original position reasoning leads to the endorsement of the difference principle. But there are two important problems with the response. To begin with, it seems unstable. After all, this response from the least advantaged concedes that, setting aside the sense of indignation, all the rest of one's interests have in fact been taken into account properly and given just as much weight as everyone else's. It is the indignation itself, and the resulting withdrawal and cynicism, that are taken to be so bad that one would not risk their occurrence. But once one recognizes that it is only one's indignation that makes one's situation intolerable, and that one's indignation is grounded solely in the appearance that one's interests have been neglected, when in fact they have been given the same weight as everyone else's, it seems difficult to maintain that indignation.

Nonetheless, indignation can be recalcitrant, and perhaps the indignation, and resulting withdrawal and cynicism, of the least advantaged will persist. Even if that is right, however, it does not yet ground the assertion of a genuine value discontinuity here. I do not doubt that suffering through indignation, withdrawal and cynicism is bad. This is true regardless of whether the indignation is grounded in a genuine disregard of one's other interests or not. And so in evaluating outcomes, it makes sense for an expectation that one will experience this indignation to negatively impact one's assessment of the outcome, perhaps significantly. But is there any reason to think that experiencing indignation of the sort Rawls imagines is discontinuously

https://doi.org/10.1017/S0953820815000278 Published online by Cambridge University Press

worse than not experiencing it? It seems not. After all, there are many activities that run the risk of inducing indignation and cynicism. I suspect that studying politics or political philosophy involves such a risk, and that actually engaging in politics or political activism does so as well, probably to a greater extent. Perhaps there is room for debate about this, but it does not really matter whether I am correct. All that matters is that a prospective student of politics or a prospective practitioner of politics, call her Chelsea, might reasonably come to think that her proposed activity or course of study might lead to indignation, withdrawal and cynicism. We can then ask whether it makes sense for Chelsea to take that as a decisive reason against the activity. Or, to home in more narrowly on the relevant issues, we can ask whether an adviser who has only Chelsea's interests in mind, and who shares Chelsea's belief that there is a risk of indignation and cynicism, should take that risk to settle the issue decisively, regardless of how small the risk is judged to be, and regardless of how much Chelsea might benefit from the activity if the indignation and cynicism do not manifest.

There is, admittedly, a coherent normative outlook according to which the answer to this question is 'yes'. By my lights, however, it is not a very appealing one. If the risk of indignation, withdrawal and cynicism were sufficiently high, I can understand why one might advise Chelsea to avoid an activity that would incur such a risk. But it seems to me to be clearly bad advice to avoid this sort of risk regardless of how slim the odds are of the undesirable outcome. I am optimistically hopeful that others judge similarly, and my aim is to point out that sharing in this judgement involves denying the value discontinuity Rawls needs in order to defend the difference principle. Moreover, I take it that risks of indignation, withdrawal and cynicism crop up with some regularity in ordinary lives. They may not be everyday risks, like the risk of getting hit by a car when crossing the street, but they are nonetheless ones that occasionally factor into deliberation about large-scale life decisions. Positing the value discontinuity needed by Rawls would impact these decisions in ways that are systematically counter-intuitive. These counter-intuitive conclusions are, admittedly, not enough to underwrite a fully decisive argument against a value discontinuity here, but in so far as judgements that presuppose continuity are widely shared and difficult to uproot, a Rawlsian assertion of the discontinuity needed to defend the difference principle will be unconvincing.

What about the points of discontinuity needed to vindicate the other elements of Rawls's view? In each case, to answer this question one must examine Rawls's analysis to see if it is sufficient to support a discontinuity claim. In my view, the sort of problem Rawls encounters in the case of the difference principle recurs, although it becomes

https://doi.org/10.1017/S0953820815000278 Published online by Cambridge University Press

increasingly difficult to reach a clear conclusion for or against the relevant value discontinuity as one moves up the hierarchy of Rawls's principles. It is beyond the scope of this article to resolve these matters fully, but let me briefly consider what I take to be the most plausible of these value discontinuities, the one needed to support Rawls's first principle of justice, in order to identify what is at issue.

A discontinuity-grounded defence of Rawls's first principle of justice, regarding basic liberties for all, would require asserting that for one who possesses these liberties it never makes sense to risk them in order to make one's life better, no matter how small the risk, provided that the better life being sought is continuous in value with one's current life. This does not mean that it is never appropriate to risk sacrificing some of one's basic liberties to secure others, nor does it mean that it is never appropriate to risk sacrificing one's basic liberties for some greater cause, such as the extension of those liberties to others. It simply means that one's basic liberties are never worth risking in order to achieve (continuous) gains for oneself. Similarly, discontinuity here indicates that, again setting aside important questions about how one's actions impact others, those who lack basic liberties should be willing to give up everything they do have in order to secure those liberties for themselves.

Notice that Rawls's own explanation of why it makes sense to endorse his first principle of justice from within the original position involves understanding the basic liberties it ensures in this sort of way. He begins with the idea that persons have fundamental interests in (1) being able to understand and be motivated by fair principles of justice and (2) being able to form and revise their own conceptions of the good.[49] On my line of analysis, this amounts to the claim that these capacities, which Rawls calls the two moral powers, are so important that having them is discontinuously better than not having them.[50] Rawls then argues that certain basic liberties are necessary to ensure these capacities, and those are, by design, precisely the liberties that are guaranteed by his first principle of justice.[51] In particular, on his view,

[49] Rawls, *Justice as Fairness*, pp. 18–19.

[50] There is another way of interpreting Rawls's argument, which involves focusing on the idea that the parties in the original position have a 'higher-order' interest in securing the two moral powers, as asserted on Rawls, *Justice as Fairness*, pp. 175 and 177, and grounding the preference for Rawls's principles of justice in that higher-order interest. As I see it, though, if the assertion of a higher-order interest is interpreted as a stipulation built into the original position, then it looks much like the thickening of the veil discussed in sect. II. It is hard to see why we should adopt that original position model over one that lacks such a stipulation. On the interpretation I prefer, the idea of a higher-order interest in securing the two moral powers is itself grounded, if at all, in a value discontinuity between possessing those moral powers and not possessing them.

[51] Rawls, *Justice as Fairness*, pp. 112–13.

https://doi.org/10.1017/S0953820815000278 Published online by Cambridge University Press

freedom of thought and the political liberties (to vote, hold office and otherwise participate in politics) are necessary for the first moral power, and liberty of conscience and freedom of association are necessary for the second. These liberties can only be ensured, he claims, through the additional liberties covered by the principle: liberties of physical and psychological integrity and the liberties involved in the rule of law.

The key question, therefore, is whether having the two moral powers is discontinuously better than not having them. For the first moral power, which Rawls calls the sense of justice, I think the discontinuity claim leads to excessively counter-intuitive conclusions, comparable to those that follow from the discontinuity needed to support the difference principle. Or at least, that is the case if we accept Rawls's claim that in order to preserve one's sense of justice one needs what he calls the 'fair value' of one's political liberties, which means that everyone must have equal abilities to run for office or affect the outcomes of an election.[52] After all, as a citizen of the United States I am fairly certain that I currently lack the fair value of my political liberties. I find that regrettable, and I believe that one of the deepest problems with the political system in the United States is the concentration of political influence among the wealthy, but I do not think a reasonable self-interested calculation would tell me to sacrifice everything (apart from other liberties) in order to have a chance at achieving the fair value of my political liberties, which is what a discontinuity claim here would imply.

With respect to the second moral power, however, I think it is far less clear that there is no relevant value discontinuity and I find it at least plausible to suppose that there is. In other words, it seems to me that Rawls may very well be right that it would be foolish to incur even a slight risk to the liberties necessary for one's ability to form and revise a conception of the good for the sake of even a substantial gain elsewhere. And the plausibility of this claim is significantly enhanced by Rawls's revision, between *A Theory of Justice* and *Justice as Fairness*, from having the first principle guarantee 'the most extensive total system' possible of equal basic liberties to 'a fully adequate scheme' of them.[53] With that more moderate conception of basic liberties in place, the strategy of undermining the plausibility of the relevant value discontinuity by finding strikingly counter-intuitive claims that follow from it is far more difficult to pursue than in the case of the difference principle.

[52] Rawls, *Justice as Fairness*, pp. 148–50.
[53] Rawls, *Theory*, p. 266 and Rawls, *Justice as Fairness*, p. 42.

https://doi.org/10.1017/S0953820815000278 Published online by Cambridge University Press

Harsanyi offers some examples that are potentially relevant here, but none that are fully compelling. In one instance, he calls attention to voluntary agreements in which one gives up certain freedoms for economic compensation, and to public policies that restrict all citizens' freedoms in order to promote societal economic interests.[54] Harsanyi does not identify the agreements or policies he has in mind, but for any more specific example it is open to Rawls to contend either that the trade-offs are entirely above the point of discontinuity, which is to say that all parties retain the fully adequate scheme of liberties protected by the first principle, or that the trade-offs are ill advised. Perhaps one can imagine a specific case that avoids both horns of this dilemma, but it is difficult to do so in a way that saddles defenders of discontinuity with a clearly, and objectionably, counter-intuitive conclusion. Elsewhere, Harsanyi claims that economic growth in underdeveloped countries may require some restrictions on civil liberties, and that such curtailment may therefore be appropriate.[55] But this example has similar difficulties to the previous one. As Harsanyi admits, it is not at all clear that economic growth could justify substantial or complete suppression of civil liberties, which suggests that even if he is right about the justification of limited infringement, such infringement may not push anyone below the 'adequate scheme' protected by Rawls's first principle, and so may be consistent with the value discontinuity Rawls needs.

I think a more promising avenue, in the attempt to undermine the plausibility of this value discontinuity, is to focus on activities that generate risks of kidnapping and enslavement, such as tourist travel to dangerous areas. In this sort of case it is perfectly clear that the negative outcome being risked involves the loss of one's basic liberties, even on a narrow conception of those liberties. The trouble here is that it is not entirely obvious that such travel, if done purely for self-interested reasons, is sensible. It may instead be the case that people who run these risks do so foolishly, out of wishful thinking, or an unjustified sense that 'it could never happen to me'. My own view is that although this may be an accurate description of many real-world cases, when the risk is sufficiently small it may very well be sensible to incur it. Rawls, in asserting the value discontinuity needed to support his first principle of justice, is committed to my being wrong about this. I therefore think he is mistaken, but I also concede that my view of this is contestable, and I do not think I can generate a powerful objection to Rawls on this basis. In other words, unlike the case of the discontinuity needed

[54] Harsanyi, 'John Rawls' Theory of Justice', p. 73.
[55] Harsanyi, 'Can the Maximin Principle', p. 602.

https://doi.org/10.1017/S0953820815000278 Published online by Cambridge University Press

to support the difference principle, I do not think I have identified, or can identify, an evidently counter-intuitive consequence of positing discontinuity here.

VIII. CONCLUSION

Hare, Harsanyi and Gibbard have repeatedly raised a line of objection to Rawls that threatens to pose a formidable challenge to his analysis. But, especially in Harsanyi's work, the challenge is formulated in a vocabulary and style that is so different from that of Rawls that the two sides of the debate often seem to talk past one another. My hope is that by focusing on Harsanyi's continuity axiom I have been able to bring the core ideas from both sides into contact without doing excessive violence to either.

One of my central claims has been that Rawls's analysis is best understood as involving the assertion of several key value discontinuities. In so far as Rawls is able to justify positing those discontinuities, he is in a position to defend his principles of justice not merely by showing that utilitarian calculations lead to their endorsement but by demonstrating that the relevant utility calculations, at least as Harsanyi sets them up, cannot even be formulated properly. As a result, assessing the debate requires determining whether Rawls is entitled to posit the relevant discontinuities.

I have attempted to show that Rawls's arguments in defence of the difference principle are not sufficient to justify positing the value discontinuity needed to vindicate that element of his view. Rather than embracing discontinuity here, I have argued that it is more appropriate, and more consistent with relatively uncontroversial everyday normative judgements, to view the negative outcomes whose avoidance Rawls cites as motivation for adopting the difference principle as genuinely bad, but nonetheless on a value continuum with other outcomes. As I have indicated, I suspect the same is true for the other points of discontinuity needed to support Rawls's principles of justice. Nonetheless, in the case of Rawls's first principle, or at least the part of that principle grounded in what Rawls calls our second moral power, I think it is far more difficult to show that acceptance of the relevant discontinuity is inconsistent with widespread, uncontroversial normative judgements. There is room, therefore, for a potentially convincing Rawlsian defence of that principle that involves blocking arguments like Harsanyi's by rejecting his continuity axiom. I have not fully resolved this contested discontinuity claim, but if my characterization of the debate is correct, and if I am right about the prospects of defending the various points of discontinuity needed to

https://doi.org/10.1017/S0953820815000278 Published online by Cambridge University Press

support Rawls's principles of justice, the lesson is that Hare, Harsanyi and Gibbard have been pursuing a line of objection that has the potential for real bite against Rawls, although it may not be able to ground as robust an argument against him as they have supposed.[56]

sdaskal@niu.edu

[56] I would like to thank Blain Neufeld, an audience at the University of Wisconsin–Milwaukee, and a particularly helpful anonymous reviewer for their feedback, comments and advice on earlier versions of this article.

https://doi.org/10.1017/S0953820815000278 Published online by Cambridge University Press