12-2015

Rational Choice and the Original Position: The (Many) Models of Rawls and Harsanyi

Gerald Gaus
University of Arizona

John Thrasher
Chapman University, thrasheriv@chapman.edu

Recommended Citation

Gaus, Gerald, and John Thrasher. "Rational Choice and the Original Position: The (Many) Models of Rawls and Harsanyi." *The Original Position*. Edited by Timothy Hinton, Cambridge University Press, 2015, pp. 39-58.

2 Rational choice and the original position: the (many) models of Rawls and Harsanyi

Gerald Gaus and John Thrasher

2.1 The original position and rational justification

2.1.1 The Fundamental Derivation Thesis

At the outset of *TJ* Rawls closely links the theory of justice to the theory of rational choice:

> one conception of justice is more reasonable than another, or justifiable with respect to it, if rational persons in the initial situation would choose its principles over those of the other for the role of justice. Conceptions of justice are to be ranked by the acceptability to persons so circumstanced. Understood in this way the question of justification is settled by working out a problem of deliberation: we have to ascertain which principles it would be rational to adopt given the contractual situation. This connects the theory of justice with the theory of rational choice. (*TJR*, p. 16)

Indeed, Rawls proclaims that "*the theory of justice is part, perhaps the most significant part, of the theory of rational choice*" (*TJR*, p. 15, emphasis added; see section 2.2.3 below). Many have refused to take this claim literally (or even seriously), by, for example, interpreting the original position analysis as a heuristic for identifying independently true moral principles (see Dworkin, "Original Position," p. 19 and Barry, *Theories*, pp. 271–82). In this chapter we take this fundamental claim of Rawls at face value. We thus shall defend:

> *The Fundamental Derivation Thesis*: the justification of a principle of justice *J* derives from the conclusion that, under conditions *C*, *J* is the rational choice of chooser(s) *P*.

On the Fundamental Derivation Thesis that *J* is the rational choice of *P* under *C* is neither *evidence* that *J* is the correct principle nor a way of us *appreciating* or *seeing* that *J* is just. Rather *J*'s justification is derivative of *J*'s status as *P*'s rational choice; moral justification derives from justification qua rational choice. Notice that we do not say that *J*'s justification is *entirely* derivative of rational choice justification, for the set of the conditions *C* under which the choice is made (including those that identify the feasible choice set) also has justificatory relevance.

The Fundamental Derivation Thesis also sets aside the interpretation of the original position as fundamentally justifying through appeal to hypothetical consent. Consent is not, strictly speaking, ever a concern for Rawls or for any original position theorist. When Rawls tell us that his "aim [is] to present a conception of justice which generalizes and carries to a higher level of abstraction the familiar theory of the social contract" (*TJR*, p. 10), we should not read him as saying that because *J would* be the object of consent by *P* under *C*, *J is* justified. At most, insofar as we might be tempted to read *any* claims about hypothetical consent into an original position argument (and we certainly need not), the relevant claim would be that the demonstration that *J would* be the object of consent by *P* under *C* shows that *J is* the rational choice of *P*, and it is this latter claim that is truly justificatory. Rawls never says that the theory of justice is part of the theory of rational consent. Henceforth we shall entirely set aside questions of consent.

2.1.2 The attraction of the Fundamental Derivation Thesis

Our aim in this chapter is not to defend the Fundamental Derivation Thesis as a philosophic commitment in theorizing about justice, but rather to show how the two most famous original position arguments – those of John Rawls and John Harsanyi – seek to follow through on it in their justifications. Nevertheless, unless the reader has at least some appreciation of the thesis's appeal, the exercise might seem pointless.

It is essential to appreciate that the Fundamental Derivation Thesis is part of a view of moral enquiry according to which models such as the original position are not epistemological, helping us to find "moral truth interpreted as fixed by a prior and independent order of objects and relations"; the original position is part of "the search for reasonable grounds for reaching agreement rooted in our conception of ourselves and our relation to society. The task is to articulate a public conception of justice that all can live with" (*CP*, p. 306). Given this, the crux of moral enquiry is to find common reasonable grounds for accepting a conception of justice. From his earliest work, however, Rawls insisted that we have conflicting ideas and intuitions about justice; arguments based on such intuitions are typically "unconvincing. They are not likely to lead to an understanding of the basis of justice" (*CP*, p. 52). Thus in the search for reasonable agreement we are led to build on a more widely shared understanding of justification – rational choice. In employing an original position argument, we suppose only that people are capable of rational choice: they know and can act on their own interests and are capable of figuring out the likely consequences of their various choices and so on. If the justification of the principles of justice derives from this prosaic, common, notion of rational choice, then we can say that the principles pass *the identification test*: actual persons, with their own interests can verify and identify with the rationality of the principles (Gauthier, *Morals*, pp. 325ff.) That is, each can see that she too would choose them under conditions of impartiality, and so can rationally identify with them. Consulting only her deliberative rationality, each sees the principles as rational. We might say that in a diverse society, the justificatory force of rational choice is our only common touchstone.

But, of course, not all rational choices constitute choice of a conception of justice. In addition to the identification test, the principles chosen in the original position must be recognizably principles *of justice*. To be principles of justice they must pass what we might call a *recognitional test* of choice in that each can confirm that the principles are chosen from an impartial moral point of view, and so qualify as bona fide principles of justice (Gauthier, *Morals*, Chapter 8).

Thus as a model of justification, the original position has two links, one to the moral point of view and the other to the point of view of actual rational individuals. Justification in the original position succeeds if the principles are chosen from a genuinely moral point of view and a rational individual can endorse them. Without the identification test, the critics of the original position would be right to see it as a complicated mechanism for generating impartial principles that do not really derive their force from rational choice (they are not what any rational agent would choose in those circumstances). Similarly, without the recognitional test, we could not be sure that the principles chosen by rational individuals in the original position were principles of justice rather than merely principles of, say, prudence or a "modus vivendi." We need to confirm that from the moral point of view, these are the rational principles of justice – that they would be chosen by an impartial legislator. Both links are essential and it is the combination of these two types of rational choice that gives original position arguments their distinctiveness and their power.

2.1.3 A social contract or an Archimedean perspective?

A long-standing question concerning Rawls's original position argument is whether the idea of a "contract" is otiose, with the real justificatory work being done by the account of a rational choice of a single chooser. Soon after the publication of *TJ*, Sidney Alexander claimed:

> The *contractarian* aspect of Rawls's device is not essential, it is even misleading. What is essential, I think, is the *choice* aspect. Whatever worthwhile principles Rawls can validly deduce from his social contract mechanism can also be deduced as the principles that a single rational man would choose, from behind the veil of ignorance, for a social system in which he was to be assigned a role after that choice. Rawls does not need the contractual aspect, as is clear from his observation that the principles would be chosen unanimously. ("Social Evaluation," p. 604)

Rawls responds to Alexander with a number of reasons why he thinks the concept of a contract is essential: (1) it "reminds" us that separateness of persons is fundamental to justice as fairness; (2) a contract "introduces publicity conditions"; and (3) "reaching a unanimous agreement without a binding vote is not the same thing as everyone's arriving at the same choice or forming the same intention" (*CP*, p. 249). None of these considerations strikes us as especially compelling; Alexander is surely correct that the critical justification is that under conditions *C* the principle *J* would be rationally chosen, and this choice can be modeled as that of a single individual, *P*. Crucially, as Rawls acknowledges, there are "no differences to negotiate" (*CP*, pp. 249 and 120), and so the choice in the original position is rational in the sense of the norms of rational individual parametric choice, not, say, non-cooperative game theory with its strategic reasoning, cooperative bargaining theory, or principles of aggregation. This is the truly fundamental point: once difference has been eliminated, the justification in the original position is via an individual principle of rational parametric choice.

We thus shall take original position arguments as necessarily seeking to reduce the choice of principles of justice to the rational parametric choice of one individual. We thus set aside an account such as Ken Binmore's *Natural Justice*, according to which there is bargaining in the original position.[1] As we shall understand it, an original position argument seeks to provide what Rawls (*CP*, p. 511) and David Gauthier have described as an "Archimedean point" for judging society:

[1] Although he does not describe himself as an original position theorist, Gauthier's *Morals* contractarian account qualifies. If space allowed, examining his view would be especially enlightening, since so few readers appreciate its strong commitment to an Archimedean moral choice.

> Archimedes supposed that given a sufficiently long lever and a place to stand, he could move the earth. We may then think of an Archimedean point as one from which a single individual may exert the force required to move or affect some object. In moral theory, the Archimedean point is that position one must occupy, if one's own decisions are to possess the moral force to govern the moral realm. From the Archimedean point one has the moral capacity to shape society. (*Morals*, p. 233)

The original position comprises such an Archimedean point, where the rational parametric choice of one individual determines the principles of justice.

2.2 The evolution of Rawls's original position

2.2.1 The early model

As does Robert Paul Wolff in his excellent analysis *Understanding Rawls*, we identify several different versions of Rawls's original position model. Rawls first advances an initial position choice model in his 1958 "Justice as Fairness" for evaluating the justice of established social practices (*CP*, p. 52). In this model of the original position, Rawls's agents are not, strictly speaking, deliberating about first principles of justice from, as it were, an atemporal "view from nowhere." As Rawls writes, "there is no question of our supposing them to come together to deliberate as to how they will set these practices up for the first time" (*CP*, p. 53). The practices are taken to already exist. Instead, the agents are deliberating about whether any of them has a "legitimate complaint" against the practices (*CP*, p. 53). In so doing, they must also develop standards for legitimate versus illegitimate complaints, and these standards shape what will ultimately be the principles of justice.

Rawls attributes a fairly narrow utility function to the choosers: individuals are mutually disinterested, evaluating the practice in question simply on the basis of whether they believe it will be to their advantage (*CP*, p. 52). This is not a claim about the nature of humans, but a model of what a certain sort of rational actor would choose. Rawls's 1958 choosers

> are rational: they know their own interests more or less accurately; they are capable of tracing out the likely consequences of adopting one practice rather than another; they are capable of adhering to a course of action once they have decided upon it; they can resist present temptation and the enticements of immediate gain; and the bare knowledge or perception of

> the difference between their condition and that of others is not, within certain limits and in itself, a source of great dissatisfaction. (*CP*, p. 52)

Rawls also assumes that they are situated in such a way that they have roughly similar abilities and needs and can benefit from cooperation (*CP*, p. 53). They normally are unable to dominate one another, thus ensuring that they have an interest in finding common principles (Wolff, *Understanding*, pp. 28–9).

That said, the precise specification of the preferences of the choosers in "Justice as Fairness" is unclear. Rawls claims that individuals will choose principles of justice for a practice where each has a conception of "legitimate claims" that it is reasonable for others to acknowledge (*CP*, p. 59). He also writes that in a just practice persons "can face one another openly and support their respective positions, should they appear questionable, by reference to principles which it is reasonable to expect each to accept" (*CP*, p. 59). All of this is somewhat vague; it leads to substantive constraints on what sorts of principles will be acceptable without fully specifying the basis of the choice in the original position. That is, the preferences of the choosers are set out as self-interested without specifying how that self-interest constrains their utility functions. As Rawls notes, this means that the subject of their agreement will be "very general indeed" (*CP*, p. 57).

The principles that they would endorse, according to Rawls, are: (1) each has an equal right to the most extensive liberty compatible with a like liberty for all and (2) inequalities are arbitrary unless it is reasonable to expect that they will work to everyone's advantage and that positions and offices are open to all (*CP*, p. 48). We can see the first principle as setting out a baseline of equality in rights or liberties and the second principle as justifying certain deviations from equality. The second principle endorses what we might call a strong Pareto condition that a move from a more equal to a less equal social state is justified if everyone is better off in the less equal state. At this point, Rawls's description of the choosers and their information sets provides no grounds for a single rational ranking of the set of mutually beneficial social states. Just states of affairs must be on the Pareto frontier, but this, in itself, does not specify a point on the Pareto frontier that is uniquely just.

This version of the choice situation does not employ any "veil of ignorance" to eliminate knowledge of an individual's capacities and interests. A veil of ignorance excludes from choosers' information sets knowledge of their identity, introducing uncertainty as to what choice is in their best interests. In the early model Rawls introduces uncertainty in another way:

> They each understand further that the principles proposed and acknowledged on this occasion are binding on all future occasions. Thus each will be wary of proposing a principle which would give him peculiar advantage, in the present circumstances, supposing it is accepted. Each person knows that he will be bound by it in future circumstances the peculiarities of which cannot be known, and which might well be such that the principles are to his disadvantage. (*CP*, p. 53)

Here the argument is that, because he is bound by the choice of the principles over an extended period of time regulating circumstances that he cannot predict, a rational person will not seek to tailor the principles so that he gains undue advantages given his present circumstances, since these may unpredictably change, and he may end up on the losing end of rigged or inegalitarian principles. Note that this argument seeks to secure some of the results that the veil of ignorance achieves in the later formulations: it requires that a person seeks principles that are what Kohlberg was later to call "reversible": one can endorse them regardless of the position one occupies under them (*Moral Development*, pp. 190–201).

Rawls is proposing here a version of maximin reasoning. He immediately explains the upshot of his principle for choice under ignorance: "The restrictions which would so arise might be thought of as those a person would keep in mind if he were designing a practice in which his enemy would assign him his place" (*CP*, p. 54). This claim is striking; it is repeated in various formulations of the original position, including "Distributive Justice" (published in 1967) (*CP*, p. 133n) and as late as *TJR* (pp. 132–3). What is so striking is that these are the only sentences in which Rawls seems tempted to introduce multi-person strategic reasoning into the original position, even though he immediately adds "the persons in the original position do not, of course, assume that their position in society is decided by a malevolent opponent" (*TJR*, p. 133).

What is going on with this unusual appeal to strategic reasoning? Recall that the definitive argument for maximin reasoning was presented by John von Neumann, who showed maximin to be the general solution to zero-sum games (*Theory of Games*, pp. 153ff.)[2] In a zero-sum, two-person game, any gain for one player implies an equal loss for the other; if we tally up all the

[2] Rawls cities this work in note 9 of "Justice as Fairness" (*CP*, p. 56) but does not direct us to specific pages. He does, though, indicate that readers should consult the Luce and Raiffa *Games* chapters on two-person cooperative games and group decision-making. Rawls always recognized that his choice situation was not, in the end, properly modeled in zero-sum terms.

gains and all the losses, the sum will always be zero. The quintessential example is a game between enemies, where one player's gain is the reverse side of the other's loss. So *if* we did think of the original position as a zero-sum game, then maximin *would be* the uncontroversial solution. It almost seems as if Rawls thinks this argument is too good not to mention, even though in *TJ* he immediately rejects it as inappropriate.

2.2.2 The middle models

Rawls's original position undergoes a series of substantial changes in what we shall call the "middle models" – from "Distributive Justice" (1967) through "Distributive Justice: Some Addenda" (1968) to *TJ* (1971). There are substantial shifts in: (1) the construction of the information sets; (2) the description of the choosers; (3) the more explicit role of maximin as a principle of rational choice; and (4) a switch in the role of maximin, from primarily an argument for the egalitarian principle, to what seems to be the main argument in favor of "the difference principle," which is itself introduced in the middle models. We briefly consider each in turn.

(1) The veil of ignorance is introduced in an effort to specify the original position as a "suitably defined initial situation" wherein "no one knows his position in society, nor even his place in the distribution of natural talents and abilities" (*CP*, p. 132). Indeed, one knows nothing of one's own personal utility function (one's ends, goals, or values). Rawls describes this restriction on the information of the choosers as justified to create a fair bargaining problem. The veil prevents "anyone from being advantaged or disadvantaged by the contingencies of social class and fortune" (*CP*, p. 132). According to Rawls, this is essential to making his contract theory, and "ethics itself," a part of "the general theory of rational choice" (*CP*, p. 132). Of course, in the general theory of rational choice, there is no restriction on the knowledge that choosers have about themselves or their fellows. Indeed, in most specifications of rational choice, agents are assumed to have full information of their circumstances. But, as we saw in the early model, Rawls is explicitly seeking to develop a theory of rational choice under radical uncertainty, where the uncertainty is intended to induce impartiality among rational choosers. The veil of ignorance is critical in allowing us to see the choice from the original position as an Archimedean point (§1.3), in which rational choice must be made from an impartial point of view.

(2) Having eliminated personal knowledge about oneself, including knowledge of one's own aims, how is anyone to make a rational choice? As Rawls

notes, rationality alone is not an adequate basis for rational choice of the principles of right, since rationality only tells us to choose more of what is preferred, not what we should prefer. It specifies means, not ends. Real persons have interests, values, and goals – what Rawls calls a "conception of the good" – that orders their rational life plans and their ends. Choosers behind the veil lack knowledge of their conception of the good; they do not know who they are, what their capacities and abilities are, and what they value and care about. As Rawls argues, while his choosers "know that they have some rational plan of life, they do not know the details of this plan, the particular ends and interests which it is calculated to promote" (*TJR*, p. 123). Without a specification of ends, however, rational choice seems either impossible or chaotic: impossible if there is simply no basis for choice, and chaotic if everyone in the original position makes spurious or random assumptions about what they would want without the veil. In the second case, once the veil is lifted, individuals from the point of view of what Rawls calls "you and me" (*PL*, p. 28) would not have reason for seeing the principles justified in the original position as having any normative force.

To solve this problem, Rawls introduces the notion of "primary goods." Primary goods are goods of which it is reasonable to assume individuals want more rather than less, regardless of whatever else they want. Thus even behind the veil, individuals know they will want to maximize their primary goods. This solves the deep concern that without some conception of the good, rational choice behind the veil would be impossible. The introduction of primary goods also considerably simplifies choice in the original position. Since everyone wants more rather than fewer primary goods, and since all have the same knowledge of their situation behind the veil, rational choice is characterized so that "unanimity is possible; the deliberations of any one person are typical of all" (*TJR*, p. 232). These elements of the original position (the veil restricting information and the thin theory of the good) make the rational choice of the "parties" truly analogous to the rational choice of one suitably constructed person, and so the choice is indeed a parametric choice of one person against a background of fixed (non-strategic) options (§1.3).

(3) The middle models witness the rise of maximin – the rule that one should choose the option whose worst outcome is better than the worst outcome of all other options – as an explicit principle of rational choice under uncertainty and, indeed, its prominent role in the overall argument. Although, as we have said, Rawls continued to employ the zero-sum strategic game imagery even into *TJ*, certainly by his 1967 essay on "Distributive Justice" he was emphasizing maximin as a general principle for choice under

uncertainty, at least under conditions of long-term commitment. Rawls believed that the uncertainty engendered by the veil of ignorance makes reliance on the maximin rule compelling.

This is not simply an idiosyncratic idea of Rawls's. In 1951 Leonard Savage noted that the minimax principle was central to the theory of choice when the actor cannot assign probabilities.[3] In "Distributive Justice," Rawls directs us to Luce and Raiffa's discussion of maximin as a principle for choice under uncertainty. However, the appeal of maximin waned in decision theory as a preferred principle for choice under uncertainty after the 1950s (e.g. McClennen, *Rationality*, pp. 25–8). Just what is the best principle for rational choice under great uncertainty is a vexed issue, but certainly as a general rule maximin seems unduly pessimistic. As Binmore quips, "only a paranoiac would find maximin attractive in general" (*Game Theory*, p. 31).

Rawls denies that his use of maximin is based on any assumptions about risk aversion (*CP*, p. 245), even though Rawls himself describes it as "conservative" (*CP*, p. 133n). In *TJ* Rawls contends that three features of the choice in the original position make plausible reliance on this "unusual" rule. (i) A distinctive feature of maximin is that it entirely discounts probabilities. In the original position, choosing social structures and your position in them, and in which you must justify this choice to others (say, your descendants), it is reasonable to be highly skeptical of any probabilistic claims. (ii) Second, "the person choosing has a conception of the good such that he cares very little, if anything, for what he might gain above the maximin stipend that he can, in fact, be sure of by following the maximin rule" (*TJR*, p. 134). And relatedly, (iii) "the rejected alternatives have outcomes that one can hardly accept. The situation involves grave risks" (*TJR*, p. 134). The point here is that the *very* special features of the choice in the original position, plus further (rather strong) assumptions about the utility functions of the parties,

[3] In von Neumann's analysis, a solution to a zero-sum game implied that maximin leads to the same choice as minimax. Suppose Row and Column are playing a zero sum game, and payoffs are stated in terms of Row's payoffs (in a zero-sum game only one player's payoffs need to be stated, since the other player's payoffs are exactly the opposite). If Row goes first, he must choose the row with the highest minimum (maximin), since he knows that Column will, from the selected Row, choose the cell which gives Column the highest payoff, which is equivalent to that which gives Row the lowest. So, like the parties in the original position, Row is concerned with nothing but the lowest payoffs in each row. If Column goes first, she will choose the column whose maximum payoff for Row is the smallest, since Column knows that once she had chosen the column, Row will choose the cell that gives him the most. Thus the equilibrium solution is one in which maximin = minimax.

renders maximin a purely rational rule under the circumstances – one that makes the most sense given the personal aims of the parties.

(4) One of the distinctive features of the middle models is that the "maximin" criterion has two distinct meanings: as a principle of rational choice and as a principle of equity ("Distributive Justice"). Although Rawls would later come to hold that the "maximining" features of the rule of rational choice and of equity (i.e. the difference principle) constitute simply a "formal resemblance" that is "misleading" (*JFR*, p. 95), they certainly seemed systematically linked in the middle models. In the early model the second principle of justice was simply a strong Pareto condition that required that justified inequalities must fall on the Pareto frontier of mutual benefit. The principle, however, does not say anything about *where* on the Pareto frontier society must settle. By the time Rawls writes "Distributive Justice," he sees this indeterminacy as a serious problem:

> There are many such [Pareto-optimal] distributions, since there are many ways of allocating commodities so that no further mutually beneficial exchange is possible. Hence the Pareto criterion, as important as it is, admittedly does not identify the best distribution, but rather a class of optimal, or efficient distributions... The criterion is at best an incomplete principle for ordering distributions. (*CP*, p. 135)

Rawls is thus committed to specifying the second principle in a way that makes a complete ordering of social alternatives possible from the point of view of justice. Here the middle models introduce another innovation: the choosers are now making their choices from the point of view of a representative member of a specific social class.

Since the veil excludes the knowledge of individual circumstances and interests that choosers could use to maximize their expected outcome in the society, Rawls argues that some particular social position needs to be selected so that the choosers can maximize the expected outcome of that representative social class. He argues that the "obvious candidate is the representative man of those who are least favored by the system of institutional inequalities" (*CP*, pp. 137–8). Applying this analysis, the agents in the original position will choose a basic structure of society as just when "the prospects of the least fortunate are as great as they can be" (*CP*, p. 138). In this formulation of the second principle we have what he now calls the "difference principle." To arrive at the difference principle, though, several more refinements or assumptions are needed in the original position. These assumptions are "chain-connection" and "close-knittedness" (*CP*, p. 139). Rawls argues that we should assume that inequalities are "chain connected," i.e. "if an inequality raises the expectation of the lowest position, it raises the expectation of all positions in between" (*CP*, p. 139). Greater expectations for low-skilled workers improve the prospects of higher-skilled workers. Relatedly, the "close-knit" assumption claims that "it is impossible to raise (or lower) the expectations of any representative man without raising (or lowering) the expectation of every other representative man" (*CP*, p. 139). These assumptions, though empirically dubious, are essential to making this version of the difference principle consistent with the earlier, Pareto version. If it is impossible to maximize the position of the least well off without at the same time harming the better off, it is not clear that even those who did not know their position in society would choose the least well off class as the representative chooser. Why not, after all, use the median class or the average representative person as the position from which to maximize, rather than the least well off? A critical reason must involve the maximin choice rule, for the parties are primarily focused on the worst outcomes. It is no wonder that (Rawls's protestations notwithstanding) for most readers of *TJ* the task of the maximin rule of choice is to justify the maximin rule of equity.

2.2.3 The final model: adieu to justice as rational choice?

In his original Dewey Lectures, in *PL*, and in *JFR*, Rawls describes the original position as a "device of representation" (*CP*, p. 308; *PL*, p. 48; *JFR*, p. 17). The specifications and, more importantly, the rationale, of the model change again. The final model is meant to provide the answer to the question "how is it possible for there to exist over time a just and stable society of free and equal citizens who still remain divided by reasonable religious, philosophical and moral doctrines?" (*PL*, p. 47). To answer this, Rawls models the justification of principles of justice as what rational *and reasonable* agents would choose if put into an original position where the diversity of their beliefs and aims was abstracted away and, again, their only goal was to maximize prospective social primary goods. In the final version a central aim is to model the normative political implications of two relevant aspects of our moral personality – our sense of justice (the reasonable) and our capacity for a conception of the good (the rational) (*PL*, p. 52).

This approach to the justification of principles of justice seems a significant departure from the middle models. In particular, the introduction of the reasonable as a feature of the representatives in the original position appears to signal that Rawls gave up on his bold claim that "the theory of justice is a part, perhaps the most significant part, of the theory of rational choice" (*TJR*, p. 15). Rawls now insists that the project of deriving the reasonable from the rational is misguided, but his renunciation is nuanced:

> From what we have just said, this [the claim that justice is a part of the theory of rational choice] is simply incorrect. What should have been said is that the account of the parties, and of their reasoning, uses the theory of rational decision, though only in an intuitive way. This theory is itself part of a political conception of justice, one that tries to give an account of reasonable principles of justice. There is no thought of deriving those principles from the concept of rationality as the sole normative concept. I believe that the text of *Theory* as a whole supports this interpretation. (*PL*, p. 53n)

Is Rawls truly renouncing the fundamental derivation claim (§1.1)? It certainly is a mistake to think that his project in *TJ* was an attempt to derive principles of justice from one normative concept, rationality; we concur that *TJ* as a whole makes this point clear. We have tried to show, however, that the Fundamental Derivation Thesis, which is central to the rational choice approach to ethics, does not claim that morality can be reduced to rational choice and nothing else: as we have stressed, the circumstances *C* are also fundamental. In the middle models, "reasonableness" is entirely modeled by the circumstances of the choice situation: they ensured that the rational choice could be *recognized* as a moral choice (*TJR*, p. 514). In the final model, although the reasonable continues to be primarily modeled by the circumstances of choice, Rawls more explicitly integrates some elements of the reasonable into the description of the choosers themselves. This in itself would be consistent with the Fundamental Derivation Thesis. However, the claim that the rational choice only enters into the final model in an "intuitive way" does suggest that, at least in Rawls's eyes, the Fundamental Derivation Thesis has been abandoned, with rational choice demoted to something more like a heuristic role.

Certainly in the evolution from the early to the final models, we can see a clear movement. In the early model *identification* – the ability of actual rational choosers to identify with the principles as clearly rational choices – looks very strong. The principles are general and abstract, and do indeed seem rational choices for cooperative schemes facing unknown futures. But Rawls clearly worried that the early model was lacking on the recognitional dimension: it was less clear that the contract was "moral". In addition, it did not provide a complete ordering of social states, something that Rawls thought important to a theory of justice. We can thus understand Rawls's later models as seeking to enhance the recognitional features of the original position – to show that it was a genuinely Archimedean point. Whether this enhancing of recognition was ultimately at the cost of identification is, of course, a fundamental worry. Gauthier thought it was: as the models develop, the chooser from the original position is so far removed from any preferences, capacities, and talents that actual individuals have little idea of what it would mean to identify with the choices of such a empty agent (*Morality*, p. 254).

2.3 Harsanyi's models

2.3.1 The original position model

There are three main differences between Rawls's middle model (which we take as the quintessential Rawlsian original position) and Harsanyi's in the way they construct the original position: (1) Harsanyi uses an expected utility decision rule rather than maximin for choice in the original position; (2) utilities are directly compared across *persons* (not classes) in Harsanyi's original position; and (3) choice in the original position selects a version of average utilitarianism rather than the two principles as the rational principle of justice. We will examine Harsanyi's original position in some detail, and then consider whether his model meets the conditions of an original position as we have laid them out in the first section.

For Harsanyi, moral choice in the original position is a species of rational choice, but rational choice over a very specific domain. Using utility theory, he argues that individuals choose rationally under uncertainty when they choose the prospects that will lead to the highest expected utility. In individual choice, these prospects and outcomes tend to be self-regarding or at least partial to a person's concerns, friends, and family ("Morality," pp. 43–4). Moral choice is distinctive because they are choosing not from their partial point of view (on the basis of their utility function) but impartially over social systems as a whole. In his most straightforward presentation Harsanyi describes a social system, say a capitalist system, of having $n = \{1,2,3\ldots\}$ possible social positions with specific utilities related to each ("Morality," pp. 45–7). These are ranked so that social position U_1 is better, from the point of view of any given member of that society, than social position U_2 and so on. Choice from the moral point of view then can be modeled so that a given individual i who does not know who she will be in that society would seek to maximize the expected value of being in that society, assuming that there is an equal probability of being any person in the society. That is, the chooser would maximize

$$\frac{1}{n}\sum_{j=1}^{n} U_j$$

or the "arithmetic mean of all individual utility levels in society" ("Morality," p. 46). Not knowing who one would be, the impartial choice would be to choose the highest expected average level of utility in the society.[4] In addition, various different kinds of societies can be compared using this method. We can, for instance, compare the expected average level of a capitalist versus a socialist society given various assumptions.

2.3.2 The axiomatic model

Harsanyi repeatedly claims ("Cardinal Utility," "Cardinal Welfare,") that his choice of the average utility principle is a result of applying standard Bayesian decision theory to choice in the original position, at one point stating that Bayesian rationality conditions, combined with a "hardly controversial" Pareto optimality condition, "entail *utilitarian ethics* as a matter of mathematical necessity" ("Morality," p. 233). But there are several essential features of Harsanyi's model of choice in the original position that reflect his aim of modeling specifically *moral* choice in the original position that deviate from standard Bayesian decision theory. To see this, it is worth looking at Harsanyi's axiomatic formulation of his model. There are four axioms that, taken together, show that genuinely moral choice is made in the original position (*Rational Behavior*, pp. 64–9):

1 rationality of moral preferences;
2 rationality of personal preferences;
3 positive relationship between the moral preferences of a person *i* and the personal preferences of all of the members of society;
4 symmetry.

The first axiom specifies that any individual choosing over social states of affairs would satisfy the standard requirements of Bayesian rationality, that is,

[4] Harsanyi recognizes that real individuals would know their actual positions in society. The critical point, he argues, is that a person's "value judgment will still qualify as a true moral judgment as long as he judges these social situations essentially in the same way as he would do *if he did not have this information*" (*Rational Behavior*, p. 50).

that person would have a complete ordering suitable for Von Neumann-Morgenstern transformation to a linear invariant cardinal utility scale. The second holds that one's individual preference ordering would meet the basic conditions of Bayesian utility theory. The third condition is a Pareto requirement for rational social choice. If all individuals in a society weakly prefer option A to option B in their individual orderings, the moral or social ordering should not rank option B over option A. Harsanyi argues that the first two conditions are merely rationality requirements while the third is a moral requirement, but one he thinks is "surely a rather non-controversial moral principle" ("Bayesian Decision," p. 226).

The fourth condition, symmetry or what Harsanyi sometimes calls "equal treatment," is more complicated. As Michael Moehler shows, the first three conditions, taken together, entail weighted utilitarianism but do not generate average utilitarianism and do not require interpersonal comparisons of utility ("Contractarian Ethics"). In this way, without the "equal treatment" condition, Harsanyi's description of the original position would generate a principle that reflected impartiality in the original position, but it would not reflect what he takes as fundamental to morality: universality and, more importantly, impersonality ("Morality," pp. 39–41). The weight that individuals put on their rankings of particular social states of affairs would be reflected in the overall social utility function. With the introduction of the "equal treatment" condition, all separate utility functions are given equal weight. Individuals assign the same weights to alternative rankings of social states of affairs when those rankings are made in the same utility units. This is not only a simplifying assumption: it has the substantive upshot that the only distinguishing factor between different orderings of social states is the particular individual names that would be associated with them. Individuals choose rationally in the original position when they assume they might actually have any name associated with a particular social position. Another way to put this is that an impartial rational chooser in the original position weights all individual utility functions equally. As we will see this is a key assumption for generating the equiprobability assumption that generates the average rule utilitarian conclusion.

2.3.3 Equiprobability and extended preferences

Harsanyi's proof of the rationality of average utilitarianism relies on two controversial assumptions: (1) the equiprobability assumption; and (2) interpersonal comparisons of utility or what he calls "extended preferences."

The first assumes that individuals will assign equal probabilities, when they are choosing in the original position, that they will occupy any social position in the social world they are selecting. If one is evaluating a capitalist society, for instance, one should place equal probability on person *i*, who is extremely poor, and person *j*, who is extremely rich. If there are *n* individuals in a society, a person in the original position will assign a probability of $1/n$ that she will be any particular person in that society (*Rational Behavior*, pp. 49–50). One chooses as if one doesn't know whether one will be Alf the teacher, Betty the factory owner, Charlie the laborer, or poor indigent Doris; one assigns equal probability to each option.

Harsanyi assumes that since the choice is made under uncertainty, that is, with no objective probability of being one person rather than another, the correct principle of choice is to assign equal probabilities to all possibilities. This is an application of Laplace's principle of indifference or insufficient reason, that if one doesn't have any reason for assuming a particular probability of one outcome over another occurring, one should assign equal probability to each outcome. As Binmore and others have pointed out, however, this doctrine of indifference is more controversial and ambiguous than it may initially seem (*Rational Decisions*, p. 128). If there are three horses running a race and one has no basis for judging their ability to win, Binmore asks, should one give the chance of a particular horse winning the probability of ⅓ by indifference (since it is one of three horses), or should one assign it a ½ probability of winning (since the horse will either win or lose and one has no more reason to think it will win or lose) (*Rational Decisions*, p. 129)? How we apply the principle of indifference in this case determines how we should bet, but the options are very different. The principle does not tell us how to apply itself.

Indeed, Rawls argues "the parties have no basis for determining the probable nature of their society, or their place in it" (*TJR*, p. 134). We have seen that Rawls argues that the parties would not use probability calculations at all, even a principle of indifference, relying instead on maximin. Rawls insists that choice in the original position should be modeled as a choice under "complete ignorance," not uncertainty, and the parties should, therefore, seek to protect themselves and their families from the worst possibilities that might befall them if they ended up in the lowest social role. Harsanyi ridicules Rawls's argument against using probabilities in the original position, arguing that it entails either (1) that we are forced to use a decision rule like maximin which is unsuited for rational choice under uncertainty and has absurd conclusions or (2) that we are being inconsistent on the basis of standard

Bayesian theories of rationality ("Can the Maximin?"). Harsanyi is quite right to point out that in standard Bayesian decision theory, subjective priors can be assigned to any gamble. The problem, as we have seen, is that the principle of indifference does not unambiguously lead to the equiprobability assumption. This does not show that Rawls is correct to object to Harsanyi's formulation of choice in the original position, only that the equiprobability assumption is not a mere extension of Bayesian decision theory. In a later essay, Harsanyi admits as much:

> In its traditional form, the principle of indifference, also called the principle of insufficient reason, asserts that when the conditions permit two or more different outcomes, yet we have no evidence favoring any particular outcome over any other, we ought to assign the same probability to each outcome. In this form, the principle is much too vague to be of any real use. It is also open to the important logical objection that it attempts to draw a positive conclusion (that of equal probabilities) from mere ignorance (from absence of information favoring any specific outcome) – an attempt that cannot possibly succeed. ("Objective Probabilities," p. 352)

Equiprobability is not an uncontroversial extension of Bayesian decision theory: it is a substantive moral assumption that makes choice in the original position impersonal as well as impartial ("Contractarian Ethics"). Harsanyi takes both of these conditions to be essential components of the moral point of view ("Bayesian Decision," p. 227).

In addition to equiprobability, Harsanyi's original position model also assumes interpersonal comparisons of utility. As with the equiprobability assumption, this is part of the "equal treatment" or symmetry assumption in the axiomatic model (section 2.3.2). Harsanyi justifies reliance on interpersonal utility comparisons as an application of what he calls "imaginative empathy" ("Morality," p. 50). We empathize with others by imagining ourselves in their position and thinking how we would evaluate a state of affairs if we were them:

> We imagine ourselves to be in the shoes of another person, and ask ourselves the question, "if I were now really in *his* position, and had *his* tastes, *his* education, *his* social background, *his* cultural values, and *his* psychological make-up, then what would now be *my* preferences between various alternatives, and how much satisfaction or dissatisfaction would *I* derive from any given alternative? ("Morality," p. 50)

Harsanyi argues that this imaginative empathy is not philosophically problematic for two reasons. First, human beings are so alike in psychology and so similarly situated that it is reasonable to think that we can actually empathize with others in this way. Second, he contends that in any case we must make interpersonal comparisons when we are making moral evaluations. Notice that Harsanyi is certainly making substantive modeling choices when he introduces equiprobability and interpersonal comparisons into his original position.

2.3.4 Identification and recognition

Does Harsanyi's original position meet the recognitional and identification conditions? Rawls denies that it meets the recognitional test, maintaining that the problem with utilitarianism and any view that makes the principles of justice the result of the rational choice of one impartial and impersonal observer is that they mistake impartiality for impersonality (*TJR*, p. 166). Impartial choice does not favor any particular point of view, but impersonal choice does not respect the separateness of persons and merges the distinct individuals of a society into one aggregative utility function. Much depends here on just what one thinks is characteristic of the moral point of view.

A more serious problem arises with the identification test. Regardless of whether impersonality is a reasonable assumption of choice from the moral point of view, it is hard to see how actual individuals could identify with the choice of an agent choosing to maximize average utility, rather than maximizing the expected utility of any particular person. As Gauthier points out, in Harsanyi's original position "the ideal actor, in maximizing expected average utility, is not maximizing expected utility – her own or that of anyone else" (*Morals*, p. 243). In moving from the rational choice of an agent maximizing her expected utility to the choice of maximizing expected average utility Harsanyi has "broken the link ensuring that each person would identify" with the chooser in the original position (*Morals*, p. 244). Put another way, actual individuals would not see the reasons of the chooser in Harsanyi's original position as reflecting their reasons. In this way, the justificatory link between the model of rational choice in the original position and the reasons of actual individuals in society would be severed. Individuals will not see Harsanyi's justification of average utility as a justification for *them*. If this is correct, Harsanyi's original position will not pass the *identification test* even if it does pass the *recognition test*. This is an independent reason for thinking that impersonality, as Harsanyi models it, is not a suitable standard for modeling moral choice in the original position.

2.4 Conclusion: rational choice and the Archimedean chooser

Why did Rawls and Harsanyi spend so much effort developing principles of rational choice for an Archimedean chooser – a chooser whose position is so constrained that his purely rational choice defines the moral realm for us all? Fundamental to any answer is a loss of faith that in a deeply pluralistic society the traditional sources of moral convictions – religion, tradition, or the moral insight of the elite – could provide the basis of a conception of morality or justice "that all can live with" (*CP*, p. 306). For this to be the case, each must be able to project himself into the Archimedean position. In doing so, he must confirm that he recognizes this as a genuine basis for moral choice and that he would choose in the way the theory indicates. As we have shown, these are difficult criteria to meet simultaneously. Both Rawls and Harsanyi sought to satisfy the recognitional criterion, providing an Archimedean position that could definitely impartially order feasible social states. In order to accomplish this, however, they developed choosers who possessed no determinate utility functions, who were stripped of capacities, interests, and aims. And, in addition, both resorted to highly contentious principles of rational choice – maximin and equiprobability. Real individuals, considering their own concerns and aims are apt to find these versions of the "moral point of view" alien: who knows what one would choose, with all knowledge of all aspects of one's individuality shorn away? We believe that Rawls's early model, in spite of its vagueness and indeterminacy, provides a more promising basis for connecting a rational Archimedean choice to real individual choices. That is a claim we shall have to vindicate elsewhere.