ON MODELING THE ORIGINAL POSITION

ERIC VON MAGNUS

Southern Methodist University

STEVEN STRASNICK, in his paper "Social Choice and the Derivation of Rawls's Difference Principle,"[1] claims to have provided a "formal model" of Rawls's original position (p. 85). The model, adopting the framework of social choice theory, consists of a "weak set of axioms" and a "judgment of priority" (p. 86). The axioms represent the "constraints on information" in the original position and a "weak notion of rationality" (p. 90). Strasnick believes the priority judgment is entailed by the initial equality assumption of the original position (p. 88-89). From the statements of this model Strasnick deduces the difference principle.

Strasnick claims that his formal procedure verifies Rawls's controversial derivation of the difference principle (which many critics have thought invalid). Since "One cannot criticize the difference principle from the context of the original position without falling into contradiction" (p. 86), he suggests that critical discussion of Rawls's theory should turn from the derivation (now proven valid) to the assumptions of the original position (p. 99). Following Nozick, Strasnick is skeptical about these assumptions, since they appear to rule out consideration of morally legitimate prior claims to the goods that are to be distributed in the original position (pp. 87-88). In overall intent, Strasnick sets the stage for using Nozick's criticisms of the original position assumptions to dispose of the difference principle.

In this paper I will (1) characterize Strasnick's formulation of the social choice problem and his use of his formal model. I will (2) develop an example in which Strasnick's social preference function and Rawls's agents in the original position would clearly make different choices, thus proving that Strasnick's model misrepresents the original position in some respect at least. I will then (3) show how this discrepancy comes about as the result of fundamental differences between Strasnick's formulation of the social choice problem—as one of finding a suitable way of aggregating individual preferences over known outcomes—and Rawls's formulation, which

Reason Papers No. 6 (Spring 1980) 25-35.

is quite different. I will show that no model using a preference aggregation framework like Strasnick's can represent all essential elements of justice as fairness. Finally I will (4) raise some questions about the Nozick-Strasnick interpretation of the original position and suggest an alternative interpretation, according to which Rawls is not vulnerable to Nozick's criticisms.

(1) Strasnick, like Arrow, formulates the problem of social choice as one of finding a suitable way of aggregating individual preferences over available alternatives to form a social preference ordering of these alternatives. (Here, alternatives are distributive states, or possible distributions of primary goods among individuals.) Strasnick modifies Arrow's formulation by allowing ordinal interpersonal comparisons of utility. Such comparisons are made by simply using a numerical index of the amount of primary goods (or income) individuals receive in some distribution as an ordinal utility index with interpersonal significance. (If Jones receives 5 units of primary goods, and Smith receives 7, then Smith's utility is greater than Jones's.) Utility comparisons are used to develop a notion of preference priority, which identifies the individuals whose preferences are weighted more heavily in deciding social preference in those cases where individual preferences conflict. Preference priorities thus make it possible for Strasnick to avoid Arrow's celebrated paradox, which is due to the unavailability in his formulation of any procedure for "handling" conflicting preferences.

In the treatment under discussion here, Strasnick assigns the highest priority to the preference of the individual who would be worst-off (whose payoff in primary goods would be lowest) if his preference were frustrated. Social preference is then identical to the preference of the "worst-off" individual, the preferences of other individuals are disregarded, and no inconsistency in the social preference ordering can occur. (Strictly speaking, social preference is that of the worst-off individual only for choices among pairs of distributions. For choices among three or more distributions a series of pairwise comparisons must be performed. The transitivity of social preference can then be used to identify the socially most preferred distribution. This will always be the distribution with the highest minimum payoff, not necessarily the distribution that would be most preferred by the individual with the lowest possible payoff.)

Social choice, then, for Strasnick involves (*a*) individual preference orderings, (*b*) judgments of preference priority, and (*c*) moving from (*a*) and (*b*) to a social preference, via some social preference function, or SPF. Given appropriate priority judgments, said to represent the initial equality assumption of the original position, and other axioms said to represent its information constraints and conception of rationality, Strasnick proves the theorem that the SPF must be the difference principle. His procedure is comparable to any use of a formal logic to test the validity of some informal argument. One must identify and paraphrase appropriately the premises and the conclusion of the informal argument. Then one must reconstruct a formal argument that proves the conclusion, given the premises. Strasnick's formal model of the original position consists simply of a set of premises that are paraphrases in his formalism of Rawls's premises, that is, Rawls's assumptions concerning freedom and equality, the veil of ignorance, and rational self-interest. Strasnick's claim to have used this formal model to verify Rawls's derivation can fail in several ways. His premises (or conclusion) may not be suitable paraphrases of Rawls's premises (or conclusion). His formal deduction may not be valid. If his deduction is valid, it is still possible that it is a different argument from the one Rawls uses, which could be invalid even though a valid argument from his premises to his conclusion exists.

One would expect that the least vulnerable part of Strasnick's procedure would be his formal deduction (though R. P. Wolff has pointed out some problems concerning its validity).[2] Raising questions concerning the appropriateness of his premises (his formal model) would seem a more promising line of attack. I will give very brief informal characterizations of Strasnick's four axioms here, even though I do not intend to criticize them, since this will provide useful detail concerning the nature of Strasnick's model. Strasnick's first axiom is Binariness and is said to make social choice a function of individual preferences and their priorities only (p. 91). (This axiom is analogous to Arrow's Independence of Irrelevant Alternatives.) His second and third axioms are Anonymity and Neutrality, which are intended to make social choice independent of the labels used to designate different individuals or alternative distributions (pp. 91-92). These three axioms are said (rather incredibly) to represent the information constraints of the original position. The fourth axiom, Unanimity, imposes a consistency requirement on the SPF. For example, if X is preferred to Y in a subsociety consisting of Smith and Jones, and similarly in a subsociety consisting of Riley and O'Brien, then X is preferred to Y in the society consisting of all four. This axiom is said to capture "an element of rationality" that is part of "the significance of the original position" (p. 93). It would seem that these axioms are extremely questionable as paraphrases of the information and rationality premises (assumptions) in Rawls's derivation argument, though I do not pursue this here.

Strasnick claims that his four axioms are consistent with most major theories of distributive justice, including utilitarianism (p. 90). It is addition of the Priority Principle to the model that renders it inconsistent with any SPF other than the difference principle. My criticism of Strasnick's model will be restricted to his Priority Principle (though in my third section I will criticize the approach to the problem of social choice embodied in his axioms).

(2) Strasnick arrives at his Priority Principle through analysis of a kind of choice situation. "In the initial situation of equality, individuals i and j will each possess the same amount of primary goods (see 62). Suppose we can increase the allotment of primary goods for one individual by transporting him to another state. If we place individual j in state u, he will receive a higher allocation of primary goods than would individual i if he were placed in state x. Since only one of these individuals may benefit, we must decide whose preference for the new state is to have greater priority." (P. 88)

I believe the following payoff matrix exemplifies the kind of situation Strasnick has in mind. Let individual i be Smith, individual j be Jones, and e be the initial situation of equality:

	Smith	Jones
e	5	5
x	6	5
u	5	7

Smith prefers x to e and is indifferent between e and u. Jones prefers u to e and is indifferent between e and x. Treating amounts of primary goods as an ordinal utility index with interpersonal significance, Smith and Jones have the same utility in e, Smith has more utility than Jones in x, Jones has more utility than Smith in u, and Jones in u has more utility than Smith in x. Total utility is greatest in u, less great in x, and least in e.

Strasnick points out that a utilitarian would assign greater priority to Jones's preference, since his gain in moving to his preferred state would be greater than Smith's gain in moving to his preferred state. Hence u would be the socially preferred state (in this two-person case) using a utilitarian SPF, since total utility is greatest in u. (P. 89) What priority judgment would be required by the original position assumptions? According to Strasnick, assigning higher priority to Jones's preference would "involve denial of a necessary property of primary goods. . . that all individuals have the same claim to them. . . . if we were correct in according j's preference a greater priority from a moral point of view, that would entail that j was entitled to more primary goods than i" (p. 89). So the preferences of Smith and Jones must be assigned the same priority in such a case. Otherwise the initial equality assumption of the original position is violated.

Strasnick formalizes this conclusion in his Priority Principle, which reads: "For all i, j, x, y, u, z, if $y_i = z_j$, then $xP_i \sim uP_jz$" (p. 89). Here, y_i is i's payoff in state y, and z_j is z's payoff in state z. The Principle says that in a case where individuals i and j receive the same payoffs in their less-preferred states, their preferences for their more-preferred states must be assigned the same priority. The symbols "$xP_iy \sim uP_jz$" are read "i's preference for x over y has the same priority as j's preference for u over z." The case represented by my matrix is obtained by letting both y and z be the same state, e, initial equality, where both individuals receive 5 units of primary goods.

Thus Strasnick's formulation assigns the same priority to Jones's preference for u over e as it does to Smith's preference for x over e. Strasnick points out later that in the two-person case the Priority Principle becomes the SPF, "a special case of the difference principle. . . . if two persons with conflicting preferences would be left equally badly-off if their preferences were frustrated, the social preference must be indifferent between them" (p. 98). So in the two-person case of my example, social choice is indifferent between x and u, that is, between the distributions (6, 5) and (5, 7).

How would agents in the original position, in Rawls's own formulation, view the choice between x and u? This question concerns specific properties of the choice situation defined by the original position assumptions. It is different from the question of how Rawls's difference principle would choose between x and u. According to the rational self-interest assumption of the original position, agents in it will attempt to identify the distributive arrangements in which their overall prospects are best, given that they do not know what their position will be in any of the distributions under consideration. They will choose the arrangements in which their overall prospects seem best to them. In this example there is no difficulty identifying the arrangement that offers the best prospects, since (5, 7) is obviously better than (6, 5): the two distributions have the same minimum, (5, 7) has a higher maximum, and no distinctions can be made concerning the probabilities of being in either of the two positions—an extremely easy case to decide, it would appear, given the original position choice assumptions.

So in this example, agents in Rawls's original position would prefer (5, 7) to (6, 5), while Strasnick's formal model entails that social preference will be indifferent. Since the two formulations make different choices in at least some cases, neither can be a model of the other.[3] This proves that Strasnick's claim to have provided a formal model of the original position is mistaken.

(We may note that Strasnick, in developing his Priority Principle, uses information about the payoffs individuals will receive—that is, their places in some distribution—and information concerning what their payoff will be in some second distribution, given what it was in some prior distribution. None of this is admissible in Rawls's original position.)

(3) Let us consider how the above discrepancy arises. In Strasnick's formulation, individuals form preferences over distributive states on the basis of known payoffs they will receive in each state. Individuals simply prefer more to less. Then a priority judgment is invoked, which identifies the individual whose preference is to be decisive. A social preference follows, with no danger of inconsistency, since other possibly conflicting preferences are ruled out of consideration.

In Rawls's formulation it is also true that only one preference decides the matter, and problems resulting from conflicting preferences are thereby avoided. But Rawls's ruling preference is arrived at in an entirely different way, without invoking anything resembling a notion of preference priority. Agents in the original position are asked to form preferences over entire distributions of payoffs, on the assumption that they do not know what their position will be in any distribution. They must find some rational procedure for comparing their overall prospects under one distribution with their overall prospects under another: how does one weigh possible gains and losses under one distribution against possible gains and losses under an alternative distribution? Since the agents will have to consider the possibility of being in any position, high or low, they will in effect have to take into account the interests of every person in their choice of distributive principles. For this reason, the original position choice can be viewed as a device for giving appropriate weight to the potentially conflicting interests of every person affected by the choice. Claims that principles chosen in such circumstances would be fair depend on the adequacy of this device (as much as on the elimination of biased choices by means of information restrictions). Thus the task facing agents in the original position of forming rational preferences over entire distributions of payoffs is an essential element of justice as fairness. It is entirely different from trying to establish preference priorities or deciding whose preference is to rule.

Strasnick's formulation cannot yield a social preference for (5, 7) over (6, 5), or vice versa, because suitable grounds for assigning higher priority to either individual's preference cannot be found, given his interpretation of the initial equality assumption. For Rawls, (5, 7) can be chosen over (6, 5), with no need for an account of why some person's preference should be given greater priority, since the choice is made on an entirely different rationale: when the two distributions are considered in their entirety, overall prospects are clearly better in (5, 7) than in (6, 5). If the task of forming rational preferences over entire distributions is essential to the Rawlsian conception of justice as fairness, then no social choice formulation that employs individual preferences based on known payoffs can possibly provide a model of it, since the essential task of forming such complex preferences cannot be represented within that kind of social choice framework. The example of the preceding section is thus symptomatic of a fundamental difference in approach to the problem of social choice. Rawls makes no attempt to aggregate individual preferences nor to form priority judgments. Strasnick makes no attempt to form preferences over entire distributions of payoffs.

Some additional remarks may bring out this important difference more clearly. One hard part of the choice problems facing agents in the original position is to solve the problem of how rational preferences over entire distributions of payoffs are to be formed. (A second hard part of the choice concerns estimating, as closely as possible given admissible information, the distributions likely to result under the various principles of justice being considered, but this does not concern us here.) The assumptions of the original position are not decisive concerning a proper method for forming such complex preferences. Rawls argues (rather than assumes) that maximin is the proper method, at least for the peculiar features of the choice of principles of justice in the original position. (He does not defend maximin as a general method for forming such preferences.) If maximin is adopted, then the difference principle, in some form, is chosen. Harsanyi argues, contrary to Rawls, that agents in the original position should maximize expected utility (and he claims to have the weight of Bayesian decision theory behind his argument). If his method is adopted, then the principle chosen is a kind of average utilitarianism (though this principle differs from classical principles in important ways, because of Harsanyi's employment of von Neumann-Morgenstern utilities in the choice).[4] Hare, contrary to both Rawls and Harsanyi, argues for a conservative "insurance" strategy, which assures a decent minimum income (unlike average utilitarianism) but not the highest possible, since the latter could result in excessive losses at the higher end of the distribution in exchange for small gains at the lower end, thus worsening overall prospects.[5]

The original position assumptions thus seem to allow for considerable argument about which method is most appropriate. Its direct assumptions do not obviously rule out any of the methods mentioned above. But for Strasnick, any method that does not result in the choice of the difference principle must conflict with the initial equality assumption (when conjoined with the other axioms). Yet it is hard to see where any such contradiction actually arises, and (so far as I know) no defender of maximin against critics like Harsanyi and Hare has attempted to show that these critics' choice strategies are logically incompatible with the initial equality assumption of the original position.

In Strasnick's formalism it seems impossible to even represent the essential Rawlsian task of forming rational individual preferences over entire distributions of payoffs. Consequently, he can find little sense in the critical debate concerning this aspect of Rawls's derivation (for Strasnick, such criticisms of Rawls are self-contradictory). If arguments like those of Harsanyi or Hare are even consistent—that maximin is not appropriate, that some principle other than the difference principle could be chosen, and so on—then they provide another display of the inappropriateness of Strasnick's model.

(4) Now I would like to examine the Nozick-Strasnick interpretation of the original position, particularly the initial equality assumption that gave rise to the Priority Principle and the immediate discrepancy between Strasnick's SPF and the original

position pointed out in my example. On this interpretation the choice seems to be of a rather concrete sort, concerning the fair division of some preexisting stock of goods. As Nozick puts it (quoted by Strasnick): "Imagine a social pie somehow appearing so that *no one* has any claim on any portion of it, no one has more of a claim than any other person" (p. 88). This interpretation does not allow for any functional dependence of the amount of goods available for distribution on the way such goods are to be distributed. It also does not allow for morally legitimate prior claims to a part of these goods, or to an unequal amount of them, based on an individual's role in producing them (one of Nozick's objections to Rawls).[6]

In Strasnick's words, "We must assume that all individuals have an equal claim to these goods in the initial situation, or none at all" (p. 88). On this interpretation I do not see how the kind of cases Strasnick discusses to develop his Priority Principle could come up for serious consideration. The social pie simply appears, and there is no dependence of its size on how it is distributed. But then only strict equality can be considered as a distributive policy. If 10 units were available in e, and u suddenly became an option with 12 units, why not distribute them (6, 6)? The amount of goods available will not be lessened by an equal distribution, so why even consider (5, 7)?

Furthermore, Strasnick rules out preference of (5, 7) over (6, 5) on the grounds that this grants one individual an entitlement to a greater-than-equal share of primary goods, in violation of the equality assumption. But shouldn't preference for (6, 7) over (5, 5) be ruled out on the same grounds? Surely this grants unequal entitlements. Rawls's difference principle would, of course, sanction such a preference, and Strasnick's SPF would also. But apparently this is not consistent with the interpretation of the initial equality assumption that led to the Priority Principle. This interpretation should rule out any distribution other than strict equality. And since the amount of goods available for distribution will not be affected by the way they are distributed, there is no need to consider any unequal distributions in the first place. ("individuals have an equal claim . . . or none at all.")

It is clear, however, that Rawls assumes a functional dependence of the amount available for distribution on the way it is distributed. He also allows claims to unequal parts of the social pie, based on roles in producing the pie. Larger incomes are viewed as incentives to greater production (or as a means of achieving an efficient allocation of labor) and are justified when they contribute (maximally) to the welfare of the lowest station. If, following a suggestion of Nozick's, we were to start with a Rawlsian just distribution and interchange persons in income stations in such a way as to maintain the same distributive pattern, or the same minimum income—assuming this were even possible—the resulting new distribution would not necessarily be just on the Rawlsian test, because income difference would no longer be tied to and justified by their contribution to the welfare of the lower positions.[7] These overt features of Rawls's theory are plainly incompatible with Nozick's interpretation of the original position choice, as one concerning the distribution of a preexisting stock of goods, without consideration of claims or unequal entitlements based on productive roles.

A variety of considerations seem, then, to call for an alternative to the Nozick-Strasnick interpretation of the original position choice. In speaking of morally legitimate claims to parts of the social pie that are prior to the original position, Nozick clearly presupposes some kind of more fundamental normative structure upon which such claims are based. Rawls's use of the original position choice, however, seems to be directed at the most fundamental normative questions possible and thus does not allow for any prior claims of the kind Nozick mentions. It is intended to establish the most basic normative structures within which all kinds of claims arise, including those based on productive roles. The choice should not be interpreted as concerning anything as concrete as the distribution of a fixed stock of goods. Rather, it should be interpreted as establishing a basic structure within which claims arise, including those stressed by Nozick. It seems that the claims due to productive role that are recognized in the Rawlsian basic structure are more restricted than Nozick believes just. But it is not true that the original position ignores them. On the contrary, it attempts to provide a theory of their basis.[8]

If we interpret the original position choice as here suggested, the objections of Nozick and Strasnick no longer apply. There are no morally legitimate claims prior to the original position, and the problem is no longer the fair division of a preexisting fixed stock of goods—a formulation that gave rise to Strasnick's problematic, apparently incoherent, interpretation of the initial equality assumption.

I have shown by example that Strasnick's SPF and Rawls's original position yield different choices in at least some cases, thus proving that Strasnick's four axioms and Priority Principle are not a correct formal model of the original position. This discrepancy was traced to fundamental differences between Strasnick's formulation of the social choice problem—as one of finding a suitable way of aggregating individual preferences over known payoffs—and Rawls's formulation—which requires the formation of rational individual preferences over entire distributions of payoffs. This is an essential element of justice as fairness that cannot even be represented in a social choice framework like Strasnick's (an "impossibility" result of some generality). I have offered a number of criticisms of the Nozick-Strasnick interpretation of the original position choice, particularly its initial equality assumption. Some raise questions about the coherence of Strasnick's argument for the Priority Principle; others raise questions about viewing the choice as one concerning the fair division of a preexisting stock of goods. Finally, I suggested an alternative interpretation of the original position choice—as establishing the framework within which various claims may arise—which is more consistent with overt features of Rawls's theory and which is not vulnerable to the criticism that the original position assumptions rule out recognition of morally legitimate prior claims to the goods being distributed.

We may understand Strasnick's article overall as an attempt (*a*) to shift critical scrutiny of Rawls's theory away from the derivation, on the grounds that the formal model has proven it valid, and to the original position assumptions; and then (*b*) to suggest that these are vulnerable to criticisms like Nozick's, thereby disposing of the difference principle. If my analysis is correct, Strasnick has failed seriously on both points.

[1] *Journal of Philosophy* 73 (1976), no. 4, pp. 85-99. All parenthetical page references in the text are to this paper.

[2] Robert Paul Wolff, "On Strasnick's 'Derivation' of Rawls's 'Difference Principle,' " *Journal of Philosophy* 73 (1976), no. 21, pp. 849-58.

[3] A similar example and related point can be found in Alan H. Goldman, "Rawls's Original Position and the Difference Principle," *Journal of Philosophy* 73 (1976), no. 21, pp. 845-49.

[4] John C. Harsanyi, "Cardinal Utility in Welfare Economics and in the Theory of Risk-Taking," *Journal of Political Economy* 42 (1953), no. 5, pp. 434-35. Also "Can the Maximin Principle Serve as a Basis for Morality?" *American Political Science Review* 69 (1975), no. 2, pp. 594-606.

[5] R. M. Hare, "Rawls's Theory of Justice," in *Reading Rawls*, ed. Norman Daniels (New York: Basic Books, 1975), pp. 104-5.

[6] Robert Nozick, *Anarchy, State, and Utopia* (New York: Basic Books, 1974), p. 198 (and elsewhere).

[7] Ibid., p. 154.

[8] For a similar claim, see Goldman, "Rawls's Original Position," p. 847.