The (Stabilized) Nash Bargaining Solution as a Principle of Distributive Justice

MICHAEL MOEHLER
Virginia Tech

It is argued that the Nash bargaining solution cannot serve as a principle of distributive justice because (i) it cannot secure stable cooperation in repeated interactions and (ii) it cannot capture our moral intuitions concerning distributive questions. In this article, I propose a solution to the first problem by amending the Nash bargaining solution so that it can maintain stable cooperation among rational bargainers. I call the resulting principle the *stabilized Nash bargaining solution*. The principle defends justice in the form 'each according to her basic needs and above this level according to her relative bargaining power'. In response to the second problem, I argue that the stabilized Nash bargaining solution can serve as a principle of distributive justice in certain situations where moral reasoning is reduced to instrumental reasoning. In particular, I argue that rational individuals would choose the stabilized Nash bargaining solution in Rawls' original position.

I. INTRODUCTION

It is argued that the Nash bargaining solution cannot serve as a principle of distributive justice because (i) it cannot secure stable cooperation in repeated interactions and (ii) it cannot capture our moral intuitions concerning distributive questions.[1] The principle 'each according to his or her relative threat advantage' is likely to lead to a collapse in the bargaining process over the cooperative surplus, because it does not guarantee that an individual receives, if her bargaining power is too low, at least a share of the gains of cooperation that allows her to satisfy her basic needs, even though the overall cooperative surplus might permit it. As a result, the bargaining game may become a game of desperation that 'unfairly' favors the stronger party to a conflict, and the Nash bargaining solution may foster distributive conflicts instead of securing stable long-term cooperation.

It is not a peculiarity of the Nash bargaining solution that, if employed as a principle of distributive justice, the bargaining process may break down in repeated interactions. Instead, this problem applies to virtually any bargaining mechanism that allows the unrestricted use of bargaining power among conflicting parties. In this article, I restrict discussion to the Nash bargaining solution because I argue,

[1] See, for example, John Rawls, *A Theory of Justice* (Cambridge, 1971/1999), p. 116, and Amartya Sen, *Collective Choice and Social Welfare* (San Francisco, 1970), p. 121.

© Cambridge University Press 2010
Utilitas Vol. 22, No. 4, December 2010
doi:10.1017/S0953820810000348

https://doi.org/10.1017/S0953820810000348 Published online by Cambridge University Press

first, that the standard Nash bargaining solution can be modified so that it can maintain stable cooperation among rational bargainers. I call this modified version of the Nash bargaining solution the *stabilized Nash bargaining solution*. The principle defends justice in the form, 'each according to her basic needs and above this level according to her relative bargaining power'.

Second, I argue that the stabilized Nash bargaining solution can serve as a principle of distributive justice in certain situations where moral reasoning is reduced to instrumental reasoning. In particular, I argue that rational individuals, who are reasonable in the sense that they accept Rawls' original position as a legitimate device to derive principles of justice, would choose the stabilized Nash bargaining solution in Rawls' original position. Rational individuals would choose the stabilized Nash bargaining solution over Rawls' difference principle and Harsanyi's average utility principle if they were placed in Rawls' original position and had to decide on a principle of distributive justice for the basic structure of society.[2]

The article is organized as follows. In section II, I describe the Rawlsian contractualist framework in which the argument is advanced, and I specify two conditions that a principle of distributive justice must fulfill from the perspective of rational individuals in order to secure stable long-term cooperation. In section III, I argue that the standard Nash bargaining solution fulfills the first condition but not the second one, and therefore must be amended to what I call the stabilized Nash bargaining solution. Section IV links the discussion to the Rawls–Harsanyi dispute, and section V presents the argument for the stabilized Nash bargaining solution in Rawls' original position. Section VI concludes with final considerations concerning the application of the stabilized Nash bargaining solution.

II. THE CONTRACTUALIST FRAMEWORK

Assume a society whose members are rational and reasonable in Rawls' specific sense. That is, the group members are not only self-interested, but they have the capacity to form, revise, and pursue a conception of the good life and to act from an effective sense of justice,[3] and they possess a 'particular form of moral sensibility that underlies the desire to engage in fair cooperation as such, and to do so on terms that

[2] For the difference principle, see Rawls, *A Theory of Justice*, pp. 65–73, and *Justice as Fairness – A Restatement*, ed. Erin Kelly (Cambridge, 2001), pp. 42–3. For the average utility principle, see John 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.

[3] See Rawls, *Political Liberalism* (New York, 1993), p. 19.

https://doi.org/10.1017/S0953820810000348 Published online by Cambridge University Press

others as equals might reasonably be expected to endorse'.[4] In short, the members of society have not only the capacity to reason rationally, but they are moral beings of a particular kind.

Let us assume that the members of this society live in a world of moderately scarce resources in which stable long-term cooperation is necessary for all group members to reach an adequate living standard, and more generally, that the members of society live under the 'circumstances of justice' described by Rawls, which include the 'fact of reasonable pluralism'.[5] The fact of reasonable pluralism entails that, although the members of society hold reasonable comprehensive doctrines that express their religious, philosophical, moral, and political views, the individuals' conceptions of the good life may be irreconcilable, and no such particular conception of the good life can be judged objectively to be morally superior to others.

In order to reach agreement on fair terms of cooperation in such a reasonably pluralistic society and to determine unambiguously principles of justice for its basic structure, Rawls introduces the *original position*.[6] The original position is an analytic device to rationally derive principles of justice that correspond to the moral sense of reasonable individuals, who embrace in particular the moral ideals of freedom, equality, and fairness. As such, Rawls designs the original position so that it reflects these moral ideals, especially by introducing the veil of ignorance. The veil ensures that no representative of the different groups of society in the original position is favored in her rational choice of basic principles of justice. In Rawls' own words, 'the *reasonable* conditions imposed on the parties in the original position constrain them in reaching a *rational* agreement on principles of justice as they try to advance the good of those they represent'.[7]

The specific moral design of the original position allows Rawls, on the one hand, to employ a rational choice decision procedure, which is supposed to yield a unique result and thus solve the moral question without ambiguities, in order to derive principles of justice for a reasonably pluralistic society and, on the other hand, to ensure that the principles derived are perceived as fair by all reasonable members of society. According to Rawls, the notion of reasonableness is modeled adequately by the moral design of the original position together with the assumption that individuals are rational in the original position.

[4] Rawls, *Political Liberalism*, p. 51.

[5] See Rawls, *A Theory of Justice*, pp. 109–12, and *Justice as Fairness – A Restatement*, pp. 84–5.

[6] See Rawls, *A Theory of Justice*, pp. 102–68.

[7] Rawls, *Justice as Fairness – A Restatement*, pp. 81–2, italics added. For simplicity, I speak only of *individuals* in the following, and not of representatives of individuals or representatives of groups of individuals in Rawls' original position.

https://doi.org/10.1017/S0953820810000348 Published online by Cambridge University Press

As a consequence, all reasonable members of society have also reasons to follow the principles of justice to be derived in the original position in the real world, if others do so, too, although their choice of principles is purely rational in the original position.

Let us assume hypothetically that the members of our society are placed in Rawls' original position and they have already found agreement on rules of cooperation concerning political rights and access to jobs and opportunities. For simplicity, let us assume that the group members agree with Rawls' liberty principle and the fair equality of opportunity principle under the veil of ignorance, and they consider now the question of distributive justice in a narrow sense.[8] The individuals now decide on a principle that regulates the distribution of the gains of cooperation in their society in order to secure beneficial stable long-term cooperation.

If we ignore for a moment the precise conditions of the original position, in particular the informational restrictions imposed on agents by the veil of ignorance, then what are the conditions that a *rational* individual, who aims to maximize her individual utility in distributive conflicts, would demand to be fulfilled by such a principle of distributive justice? In the following, I argue that a principle of distributive justice must fulfill two conditions from the perspective of a rational individual who has an overall interest in stable long-term cooperation.

First condition: A rational individual agrees only with a principle of distributive justice that does not unduly restrict her freedom of action. As such, a principle of distributive justice must allow an individual to defend her preferences in conflicts over the cooperative surplus at least as much as possible according to her actual capacities, restricted only by the individual's goal of securing stable long-term cooperation.

A rational individual would, of course, like to defend her preferences in conflicts over the cooperative surplus more than her actual capacities allow her to do. However, her minimal demand is to do so at least as much as possible according to her actual capacities, subject to the side-constraint of maintaining stable long-term cooperation. To be precise, instrumental rationality *requires* an agent to defend her preferences as much as possible according to her actual capacities in conflict situations, because only then does the agent maximize her individual utility in the particular instances. This is not a psychological assumption about human nature, but a demand of

[8] The latest statement of the liberty principle and the fair equality of opportunity principle can be found in Rawls, *Justice as Fairness – A Restatement*, p. 42. Whether the individuals choose Rawls' liberty principle and the fair equality of opportunity principle in the original position, or slightly different principles, is irrelevant for my argument that focuses exclusively on the domain of the difference principle.

https://doi.org/10.1017/S0953820810000348 Published online by Cambridge University Press

instrumental rationality, if combined with the ideal of individual utility maximization.

Second condition: A rational individual consents to a principle of distributive justice that guarantees stable long-term cooperation only if the principle grants her at least the means that she needs to maintain her life as a basis for conflict resolution, if the overall cooperative surplus permits it. For simplicity, I refer to the amount of goods that an individual needs to maintain her life as her *minimum standard of living*. An individual's minimum standard of living includes, at its most basic level, the protection of an individual's physical integrity and her existence as a separate agent.[9]

If an individual is not granted the amount of goods that secure her minimum standard of living as a starting point for resolving distributive conflicts, she loses interest in stable long-term cooperation, because future cooperation is not possible for her. As a consequence, the individual may try to disturb the stability of the cooperative framework through negative actions, such as revolution, war, and the like, by threatening the lives of others, or by the destruction of scarce resources. Such negative actions may slow down social and economic growth or they may render cooperation unbeneficial by creating high costs to deter such negative actions.

One may object that the demand that a principle of distributive justice must grant an individual at least the means that she needs to maintain her life as a *basis* for conflict resolution, if the overall cooperative surplus permits it, is too strong. It would suffice for an individual to reach this level as an outcome of the process of conflict resolution, and not as a starting point in order to have an interest in stable long-term cooperation. This condition is insufficient, however, as I now explain.

From the perspective of a rational individual, the advantages of a distributive principle, aside from not having to fear losing one's life in distributive conflicts, are primarily the additional social and economic gains that stable long-term cooperation makes possible. As such, a rational individual will fully agree with a principle of distributive justice only if she can hope to participate in these additional gains that the principle generates. This condition is not necessarily fulfilled if the principle allows an individual only to reach her minimum standard of living as a result of the process of conflict resolution.

[9] The notion of a minimum standard of living is vague and can be defined in various ways. For the purpose of this article, I assume that the members of society come to an agreement on the determinants that define an individual's minimum standard of living and on a way to measure it. I return to this point in section VI.

https://doi.org/10.1017/S0953820810000348 Published online by Cambridge University Press

Assume, for example, an individual who currently lives at her subsistence level and who is so weak that she expects to be by far the weakest individual in almost all distributive conflicts that she will face. This individual cannot expect to be able to improve her situation over time, and thus to have access to the additional gains of cooperation that the distributive principle makes possible, if the principle allows her to reach her minimum standard of living only as a result of the process of conflict resolution.

In more general terms, if an individual's ability to cooperate is limited solely to acquiring the goods that she needs to survive, and a principle of distributive justice guarantees her this subsistence level only as a result of her endeavors, then the individual does not necessarily have access to the additional gains of cooperation that the principle generates, and consequently will lose interest in the principle. For a rational individual to agree fully with a principle of distributive justice, the principal intent of which is to generate additional gains of cooperation by securing stable long-term cooperation, the individual must be assured that she can participate in these gains. To this end, the principle must allow an individual to enter the process of conflict resolution at least from her minimum standard of living, because otherwise, in the worst case, she may be cut off entirely from the additional gains.

In sum, from the perspective of a rational individual a principle of distributive justice that aims to secure stable long-term cooperation must fulfill the two minimal conditions described. In the following, I argue that the Nash bargaining solution fulfills the first of these two conditions, but not the second one.

III. THE (STABILIZED) NASH BARGAINING SOLUTION

Nash argues that rational individuals, or more precisely two rational individuals, who face a distributive conflict over a divisible good and who aim to maximize their individual utility in the process of conflict resolution, would agree with the bargaining outcome that maximizes the product of the differences between the utilities that the individuals gain at the disagreement point from which the bargaining process takes place and the cooperative outcome.[10] To support this conclusion, Nash argues that the bargaining solution must fulfill certain formal

[10] See John Nash, Jr., 'The Bargaining Problem', *Econometrica* 18 (1950), pp. 155–62, and 'Two-Person Cooperative Games', *Econometrica* 21 (1953), pp. 128–40. The individuals' preferences are assumed to be represented by von Neumann–Morgenstern utility functions. In addition, I assume that, in our specific case, the conflicting parties accept the disagreement point as a basis for conflict resolution. I return to this assumption in section VI.

https://doi.org/10.1017/S0953820810000348 Published online by Cambridge University Press

requirements, namely, the axioms of scale invariance, Pareto efficiency, symmetry, and independence of irrelevant alternatives.[11] Nash proved that if the bargaining solution fulfills these four conditions, as his bargaining solution does, then a unique bargaining outcome always exists.

Some debate exists in the literature as to whether the Nash bargaining solution is the only rational outcome of bargaining situations that do not invoke interpersonal comparisons of utility and that are non-ethical in this sense.[12] What counts as the rational bargaining outcome depends primarily on the axiomatic formulation of the bargaining theory. The first three Nash axioms are commonly regarded as uncontroversial, but there is some debate about the plausibility of the independence of irrelevant alternatives axiom. Kalai and Smorodinsky,[13] and Gauthier,[14] for example, replace this axiom by a principle of monotonicity, which leads them to a different bargaining solution. For a two-person bargaining situation, the Kalai–Smorodinsky solution and Gauthier's bargaining solution are identical.

In the following, I do not debate the plausibility of the independence of irrelevant alternatives axiom or of any of the other Nash axioms. Instead, I provide an argument as to why rational individuals would agree, if they ignore long-term considerations, to the Nash bargaining solution as an outcome in conflict situations where moral reasoning is reduced to instrumental reasoning because the parties to a conflict cannot refer to moral standards to resolve their disputes. I argue that rational individuals would do so for three reasons.

First, in conflict situations where individuals cannot rely on moral standards to resolve their disputes, instrumental rationality requires rational agents to use their unrestricted bargaining power to fulfill their preferences, because only then do they maximize their utility in the particular instances. As a result, a rational individual will accept only a bargaining outcome that assigns her a share of the

[11] For further discussion of these axioms, see R. Duncan Luce and Howard Raiffa, *Games and Decisions* (New York, 1957), pp. 126–7.

[12] The best-known bargaining solution that relies on interpersonal comparisons of utility is the egalitarian bargaining solution. See, in particular, Ehud Kalai, 'Proportional Solutions to Bargaining Situations: Interpersonal Utility Comparisons', *Econometrica* 45 (1977), pp. 1623–30. Ken Binmore also defends such a proportional bargaining solution as an ethical solution concept in the context of his social contract theory. See Binmore, *Game Theory and the Social Contract* (Cambridge, 1994/1998), in particular, vol. 2, pp. 393–5. Other explicitly ethical bargaining solutions are defended by so-called arbitration models that specify bargaining outcomes that satisfy certain moral or fairness criteria. See, for example, Richard Braithwaite, *Theory of Games as a Tool for the Moral Philosopher* (Cambridge, 1955).

[13] See Ehud Kalai and Meir Smorodinsky, 'Other Solutions to Nash's Bargaining Problem', *Econometrica* 43 (1975), pp. 513–18.

[14] See David Gauthier, *Morals by Agreement* (Oxford, 1986), pp. 113–56.

https://doi.org/10.1017/S0953820810000348 Published online by Cambridge University Press

cooperative surplus that is *proportional to her bargaining power*. This condition, compared with other bargaining solutions, is best fulfilled by the Nash bargaining solution, according to which an individual ends up with what she can expect to gain according to her actual bargaining capacities. The Nash bargaining solution ensures that each individual receives a share of the cooperative surplus that is proportional to her relative bargaining power in a particular conflict situation.

Second, the Nash bargaining solution fulfills the first of the two conditions described in the previous section. It allows the parties to a conflict to defend their preferences in the process of conflict resolution *as much as possible* according to their actual capacities, because only then will the Nash product be maximal. This condition is captured by Nash's third axiom of Pareto efficiency. The efficiency condition is usually also fulfilled by other bargaining solutions. The Kalai–Smorodinsky solution, however, does not fulfill this demand in general, but does so only in bargaining situations with two players, as Roth has shown.[15]

Third, the Nash bargaining solution is confirmed by the non-cooperative bargaining solutions of Zeuthen,[16] Harsanyi,[17] and Rubinstein.[18] That is, if one considers not only the bargaining outcome, but also the steps of the bargaining process of ideally rational individuals, then one also ends up with the outcome of the Nash bargaining solution, given plausible assumptions about the bargaining process. The Nash bargaining solution allows rational individuals to benefit as much as they would have done if they had engaged in a direct bargaining process with each other. In other words, the Nash bargaining solution is grounded securely in non-cooperative game theory, and thus represents a natural agreement point for rational individuals in conflict situations where their bargaining process is morally unrestricted.[19]

[15] See Alvin Roth, 'An Impossibility Result Concerning n-Person Bargaining Games', *International Journal of Game Theory* 8 (1979), pp. 129–32. In this context, see also H. Peyton Young, *Equity: In Theory and Practice* (Princeton, 1994), pp. 121–2, who argues that the Kalai–Smorodinsky solution is inconsistent in multi-party bargaining games.

[16] See Frederik Zeuthen, *Problems of Monopoly and Economic Warfare* (London, 1930), ch. 4.

[17] See Harsanyi, 'Approaches to the Bargaining Problem Before and After the Theory of Games: A Critical Discussion of Zeuthen's, Hicks', and Nash's Theories', *Econometrica* 24 (1956), pp. 144–57, and *Rational Behavior and Bargaining Equilibrium in Games and Social Situations* (Cambridge, 1977), pp. 141–66.

[18] See Ariel Rubinstein, 'Perfect Equilibrium in a Bargaining Model', *Econometrica* 50 (1982), pp. 97–109. To be precise, Nash himself already presented a non-cooperative approach to his axiomatically derived bargaining solution in 'Two-Person Cooperative Games'.

[19] For further support of the Nash bargaining solution in this context, see, for example, Brian Barry, *Theories of Justice* (Berkeley, 1989), pp. 22–4; Ken Binmore, 'Bargaining and

https://doi.org/10.1017/S0953820810000348 Published online by Cambridge University Press

As indicated, however, the standard Nash bargaining solution does not necessarily satisfy the second condition that a principle of distributive justice must fulfill from the perspective of rational individuals in order to secure stable long-term cooperation. The Nash bargaining solution does not guarantee that an individual receives at least a share of the cooperative surplus that is large enough for her to maintain her minimum standard of living, even though the overall cooperative surplus might permit it. In order for the Nash bargaining solution to fulfill this demand, this requirement must be added to the conditions that the bargaining outcome must fulfill. I call the resulting principle the *stabilized Nash bargaining solution*.

The difference between the standard Nash bargaining solution and the stabilized Nash bargaining solution is that, if distributive conflicts are resolved according to the latter, then the stronger party to a conflict will not always be able to exert all of her strength on her opponent. Instead, the stronger party must support her opponent, if her opponent falls below her minimum standard of living at the beginning of the process of conflict resolution, so that her opponent can enter the bargaining process at least from her minimum standard of living. The stronger party to a conflict must give up part of her potential gain in the bargaining over the cooperative surplus if her opponent is too weak to secure the means that she needs to survive as a basis for conflict resolution.

I will not discuss further the distributional differences between the standard Nash bargaining solution and the stabilized Nash bargaining solution because I think that the main distinction between the two principles is intuitively clear. However, for the reader who has a deeper interest in the precise properties of the stabilized Nash bargaining solution, a more rigorous analysis of the proposed bargaining principle is offered in the appendix. In the following, I address the question as to whether the stabilized Nash bargaining solution is an attractive principle of distributive justice for rational individuals in Rawls' original position. To this end, I link the discussion to the Rawls–Harsanyi dispute.

Morality', *Rationality, Justice and the Social Contract*, ed. David Gauthier and Robert Sugden (Ann Arbor, 1993), pp. 131–56, *Game Theory and the Social Contract*, vol. 1, pp. 80–4 and vol. 2, pp. 77–95, and *Natural Justice* (Oxford, 2005), pp. 25–7; and Brian Skyrms, *Evolution of the Social Contract* (Cambridge, 1996), p. 107. See also the later Gauthier who argues, in 'Uniting Separate Persons', in *Rationality, Justice and the Social Contract*, pp. 176–92, at pp. 176–9, that the real work of justifying his bargaining solution in the circumstances of the social contract is to show that it coincides with the Nash bargaining solution.

https://doi.org/10.1017/S0953820810000348 Published online by Cambridge University Press

IV. THE RAWLS–HARSANYI DISPUTE

Although the members of our society are assumed to be reasonable in real life and, as such, have reasons to enter the original position in their attempt to derive a principle of distributive justice, their decisions in the original position are assumed to be merely rational. In the original position, individuals are assumed to select a principle of distributive justice that best advances their own real-life prospects, whatever their life plans may be. The individuals' aim to further their individual interests in the original position is restricted only by the demands that the agents are mutually disinterested and guided neither by envy or hate, nor by a desire simply to dominate others, nor by specific attitudes towards risk in their choice of a principle of distributive justice.[20] Accepting these restrictions on the individuals' preferences in the original position, and restricting the discussion to the topic of distributive justice in a narrow sense by not considering Rawls' arguments for the liberty principle and the fair equality of opportunity principle, what is the rational outcome of the decision-making process in Rawls' original position?

Rawls argues for the difference principle. An unequal distribution of social primary goods (excluding basic liberties and job opportunities, which must be equally distributed or equally accessible to all members of society, according to Rawls), is justified only if such an unequal distribution is advantageous for each member of society, and in particular for the least advantaged. The least advantaged members of society must gain most from an unequal distribution, as compared to an equal distribution of social primary goods, because Rawls believes that rational individuals would be very cautious if they had to decide on a principle of distributive justice that significantly determines the prospects of their later lives.

In fact, Rawls argues that rational individuals would be so cautious in the original position that they would assume that they will end up in the worst-off position in society after the veil of ignorance is lifted, and instrumental rationality then requires the individuals to maximize the prospects of this position under the veil of ignorance. Rawls assumes that rational individuals would like to ensure in the original position that in any case, i.e., no matter how small the probability of being in the worst-off position in society may be, they can attain at least the amount of social primary goods that allows them to live a minimally decent life.

[20] See Rawls, *A Theory of Justice*, pp. 12, 124, and 464, and *Justice as Fairness – A Restatement*, pp. 87–8.

https://doi.org/10.1017/S0953820810000348 Published online by Cambridge University Press

The problem with this line of reasoning is that if the difference principle is in effect, an individual can be sure that she can maximize her individual utility based on her actual capacities only if she is in the worst-off position in society after the veil of ignorance is lifted. In fact, if she is in this position she receives more than she can expect to gain based on her actual capacities. By contrast, if an individual does not end up in the worst-off position, then she might constantly have to give up part of what she could have gained based on her actual capacities, because as long as not all social primary goods are equally distributed in society there is always a worst-off group. The difference principle demands that the stronger members of society support the weakest group members, in the worst case for the non-worst-off group, permanently and independently of the absolute standards of living that the weakest members of society have reached, assuming that the non-worst-off group gains at least marginally from inequality.

Many criticisms of the difference principle are well known, including Harsanyi's skeptical response to Rawls.[21] Harsanyi argues that it is irrational to assign such a high probability to being in the worst-off position of society in the original position. Instead, a rational individual would assign an equal probability to all possible outcomes in a situation where she has no information whatsoever about the likelihood of ending up in any of the positions of society and, as a consequence, a rational individual would maximize the average utility of the different positions of society. In terms of normative decision theory, Harsanyi argues that a rational individual would apply the principle of insufficient reason (the Laplace rule) in the original position, whereas Rawls argues for the maximin rule.[22]

Rawls' concern with Harsanyi's argument for the average utility principle is that the utilitarian calculus does not protect an individual from ending up in a disastrous position that does not permit her to fulfill her basic needs in real life, because the utilitarian calculus averages out gains and losses across people, and thus permits significant variance. A utilitarian response to this criticism would be to introduce a social minimum and then maximize the average utility of society above this level. But once again, Rawls disagrees with this solution.[23]

In the following, I argue that rational individuals would choose neither the difference principle nor the average utility principle in

[21] See Harsanyi, 'Can the Maximin Principle Serve as a Basis for Morality? A Critique of John Rawls's Theory'.

[22] The later Rawls tries to make his argument for the difference principle less dependent on the use of the maximin rule in the original position. I think, however, without success. For this point, see Rawls, *Justice as Fairness – A Restatement*, p. 95, and my discussion in the following section.

[23] See Rawls, *Justice as Fairness – A Restatement*, pp. 119–30.

https://doi.org/10.1017/S0953820810000348 Published online by Cambridge University Press

Rawls' original position. Instead, they would choose the stabilized Nash bargaining solution. To support this conclusion, I argue first that rational individuals in Rawls' original position would prefer the stabilized Nash bargaining solution over the difference principle, and second that they would prefer the stabilized Nash bargaining solution over the average utility principle, even if the principle is subject to a welfare floor.

V. THE MOST RATIONAL DECISION IN THE ORIGINAL POSITION

The stabilized Nash bargaining solution fulfills Rawls' primary demand on a principle of distributive justice that no individual ends up in a disastrous position in which she cannot satisfy her basic needs. The stabilized Nash bargaining solution guarantees that each member of society obtains at least the goods that she needs to fulfill her basic needs as a starting point for resolving distributive conflicts, and that she can improve her standard of living above this level according to her relative bargaining power. This condition applies, of course, only if a society is sufficiently developed socially and economically so that a redistribution of goods above the subsistence level is possible. But the same condition applies to the difference principle.[24]

The key distinction between the difference principle and the stabilized Nash bargaining solution is that the latter does not focus primarily on the worst-off group of society, but allows each member of society to maximize her individual utility as much as possible according to her actual capacities, subject to the side-constraint that all group members can maintain at least their minimum standards of living, if society can afford it. The stronger members of society do not have to permanently support the weakest members of society in the worst case for the non-worst-off group, but their share of the cooperative surplus, as determined by their relative bargaining power, is reduced only if the weakest members of society fall below their minimum standards of living at the beginning of distributive conflicts.

In other words, the stabilized Nash bargaining solution balances what Rawls calls the 'strains of commitment' more evenly in society than does the difference principle from the perspective of rational individuals in the original position. According to Rawls, the strains of commitment are an essential consideration in the original position because individuals 'will not enter into agreements they know they cannot keep, or can do so only with great difficulty'.[25] The strains of

[24] See Rawls, *Justice as Fairness – A Restatement*, pp. 130 and 162.
[25] Rawls, *A Theory of Justice*, p. 126.

https://doi.org/10.1017/S0953820810000348 Published online by Cambridge University Press

commitment must be assessed in the original position, because in real life the members of society are assumed to be reasonable, and thus they are assumed to be sufficiently morally motivated to follow *any* principle of distributive justice that is derived correctly in the original position.[26]

The stabilized Nash bargaining solution treats not only one position in society 'fairly' from the perspective of rational individuals in the original position, but all positions, and because an individual does not know in which position of society she will be after the veil of ignorance is lifted, and thus she does not know which position she must favor in order to maximize her real life prospects, a rational individual prefers the stabilized Nash bargaining solution over the difference principle in Rawls' original position. Moreover, if an individual is unlucky in the lottery of life, the stabilized Nash bargaining solution allows an individual to receive at least the means that she needs to live a minimally decent life, if society can afford it, and the individual can improve her situation above this level according to her actual capacities.

For these reasons, it is rational for individuals in Rawls' original position to choose a principle of distributive justice that considers their relative bargaining power in distributive conflicts in real life, although individuals are equally placed in Rawls' original position. That is, although individuals face a situation of choice in Rawls' original position in which each person has a veto right, it is rational for individuals in the original position to choose a distributive principle that considers their relative bargaining power in real life, within limits. The fact that individuals do not bargain with each other in Rawls' original position does not prevent them from choosing a bargaining principle as a principle of distributive justice under the veil of ignorance.

I am aware that the conclusion that rational individuals would choose a bargaining principle in Rawls' original position is as ironic as Harsanyi's utilitarian conclusion under the veil of ignorance, because Rawls' aim was to overcome these moral alternatives. But this conclusion shows just how problematic the device of the original position, as initially described by Rawls, together with the use of rational choice theory under the veil of ignorance, is for Rawls' purpose. The original position does not unambiguously support the choice of the difference principle, if it supports this choice at all.

To clarify, if the individuals in Rawls' original position are *assumed* to care only about the minimum level of goods that they receive after the veil of ignorance is lifted, and it is *assumed* that this

[26] For a more general discussion of this point, see G. A. Cohen, *Rescuing Justice and Equality* (Cambridge, 2008), pp. 27–86.

https://doi.org/10.1017/S0953820810000348 Published online by Cambridge University Press

guaranteed level is completely satisfactory, as Rawls later stipulates,[27] then the difference principle can be justified using the device of the original position and rational choice theory, albeit in a trivial manner, because these assumptions render the choice in the original position redundant.[28] Further, there may be reasons why (i) rational individuals under different circumstances, (ii) different individuals under the described circumstances, or (iii) different individuals under different circumstances may choose the difference principle, as Rawls, in my view, convincingly argues overall in *A Theory of Justice* and in his later writings. In particular, the device of the reflective equilibrium, where reasonable and not merely rational individuals are supposed to decide unveiled about principles of justice, seems to support Rawls' conclusion in favor of the difference principle.[29]

If, however, rational individuals decide on a principle of distributive justice in the original position as initially described by Rawls without referring to factors that lie outside the described decision situation, then they will not choose the difference principle, for the reasons mentioned. If the individuals were allowed to consider factors that lie outside the device of the original position, in particular moral reasons, then one of the main purposes of the analytic device, the strength of which is precisely that moral reasoning is reduced to instrumental reasoning, would be defeated. As such, once the individuals are placed hypothetically in Rawls' original position and they are covered by the veil of ignorance, no moral considerations are permitted.

I now turn to my criticism of Harsanyi's argument for the average utility principle. Harsanyi argues that rational individuals who aim to maximize their individual utility will choose a utilitarian principle of distributive justice in Rawls' original position. In general, rational individuals are not interested directly in the aggregate amount of utility that a principle of distributive justice produces. Rather, the group benefit is relevant to them only to the extent to which it affects their *individual* gains as assigned by a specific principle of distributive justice.[30] As a consequence, rational individuals neither would agree to arrange the basic institutions of society so that the overall group utility is maximized, nor would they agree to maximize

[27] See Rawls, *A Theory of Justice*, pp. 134–5, and *Justice as Fairness – A Restatement*, pp. 97–104.

[28] For an alternative decision-theoretic defense of the use of the maximin rule in Rawls' original position, see Erik Angner, 'Revisiting Rawls: *A Theory of Justice* in the Light of Levi's Theory of Decision', *Theoria* 70 (2004), pp. 3–21.

[29] For the device of the reflective equilibrium, see Rawls, *A Theory of Justice*, pp. 18–19.

[30] For a similar point, see Gauthier, 'Bargaining and Justice', *Social Philosophy and Policy* 2 (1985), pp. 29–47, at p. 44, and 'Justice as Social Choice', *Morality, Reason, and Truth – New Essays on the Foundations of Ethics*, ed. David Copp and David Zimmermann (Totowa, 1984), pp. 251–69, at p. 255.

https://doi.org/10.1017/S0953820810000348 Published online by Cambridge University Press

the utility of the aggregate average position in society. Instead, rational individuals will choose a principle of distributive justice that best allows them to maximize their individual utility in real life.

In Rawls' original position, however, rational individuals do not know in which position of society they will be, and they also do not know how likely it is that they will end up in any of the positions in society. As previously indicated, Harsanyi argues that in this situation of complete uncertainty it is rational for individuals to assume that they will end up in any of the positions in society with equal probability, and consequently rational individuals will choose a principle of distributive justice that arranges the basic institutions of society so that the average utility of the different positions in society is maximized.

However, Harsanyi's assumption to assign an equal probability to all possible outcomes in Rawls' original position is as arbitrary as any other probability assignment with regard to outcomes in Rawls' original position, because individuals in Rawls' original position cannot justify any such choice. Due to the lack of knowledge under the veil of ignorance, rational individuals cannot justify any specific probability assignments with regard to outcomes in Rawls' original position, whatever the individuals' attitudes towards risk in real life may be. This lack of knowledge forces rational individuals to be *agnostic* in Rawls' original position and, as a result, to assign neither a high or low probability nor an equal probability to being in any of the positions in society.

In other words, although the principle of insufficient reason is correctly applied to choices with symmetric outcome spaces under complete uncertainty, such as rolling dice, the use of this decision-making rule is not necessarily justified in Rawls' original position. According to Rawls, the individuals' choices in the original position must be entirely *independent* of probability assignments with regard to outcomes, which excludes also the use of the equiprobability assumption in the original position.[31] If the equiprobability assumption

[31] Rawls is explicit that there are 'no objective grounds in the initial situation for assuming that one has an equal chance of turning out to be anybody' (*A Theory of Justice*, p. 144), and that the choice in the original position is not to be guided by probability considerations. Whether this assumption is plausible is an entirely different question that is not of concern here. For a discussion of this point, see in particular Binmore, *Game Theory and the Social Contract*, vol. 1, pp. 327–9. For further discussion of the use of the principle of insufficient reason in Rawls' original position, see Rawls, *A Theory of Justice*, pp. 134–5 and pp. 144–50; Rawls, *Justice as Fairness – A Restatement*, p. 98; and Samuel Freeman, *Rawls* (New York, 2007), pp. 175–7. With regard to the use of the principle of insufficient reason, it is also interesting that Harsanyi himself argues later in, 'Bayesian Decision Theory, Subjective and Objective Probabilities, and Acceptance of Empirical Hypotheses', *Synthese* 57 (1983), pp. 341–65, at p. 352, that the principle 'is

https://doi.org/10.1017/S0953820810000348 Published online by Cambridge University Press

cannot be justified in Rawls' original position, then Harsanyi's argument for the average utility principle in the original position fails. In this case, no bridge from individual utility maximization to the utilitarian calculus exists, and the average utility principle cannot be defended based on individual utility maximizing grounds, even if the principle is subject to a welfare floor.

To clarify this point, a rational individual in Rawls' original position will choose a principle of distributive justice that best allows her to maximize her individual utility in real life from the perspective of the original position. The average utility principle, however, does not allow an individual to do so, because the principle maximizes the prospects of hypothetical 'Miss Average', and an individual in Rawls' original position has no reason to expect that she will be in any of the positions in society with an equal probability, and thus that she will be 'average' in this sense after the veil of ignorance is lifted. In real life, the average utility principle will force an individual to suspend her individual utility maximizing behavior to the extent necessary to maximize the average utility of society, because the principle assigns priority to the group level instead of the individual level, as all utilitarian principles do.

As such, it is irrational for individuals who aim to maximize their *individual* utility to choose the average utility principle in Rawls' original position. Under the restrictions imposed on agents in the original position, the only viable alternative for rational individuals who aim to maximize their individual utility under the veil of ignorance is to choose a principle of distributive justice that distributes the gains of cooperation *independently* of an individual's chance of being in any of the positions in society. The stabilized Nash bargaining solution fulfills this demand because it allows each member of society to maximize her individual utility as much as possible independent of her actual position in society, subject to the side-constraint that all group members reach at least their minimum standards of living if society can afford it.

In other words, the choice of the stabilized Nash bargaining solution in Rawls' original position does not rely on any (implicit or explicit) probability assignments with regard to outcomes, and the stabilized Nash bargaining solution does not favor any specific position in society. As a result, the choice of the stabilized Nash bargaining solution can be justified to all members of society in the original position, in

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'.

https://doi.org/10.1017/S0953820810000348 Published online by Cambridge University Press

contrast to the choice of the average utility principle, which assigns priority to the 'average' position in society, and in contrast to the difference principle, which assigns priority to the worst-off position in society.

In sum, because a rational individual in Rawls' original position does not know in which position of society she will be, and thus does not know which position she must favor in order to maximize her prospects in real life, she will choose a principle of distributive justice that allows her to maximize her individual utility as much as possible, subject to the side-constraint that all group members can maintain at least their minimum standards of living if society can afford it, whatever her actual position in society may be. The stabilized Nash bargaining solution fulfills this demand, because independent of the position in which an individual ends up after the veil of ignorance is lifted, the individual will be able to maximize her individual utility as much as possible according to her actual capacities, given her interest in stable long-term cooperation. Further, if an individual is unlucky in the lottery of life, she will receive at least the means that she needs to live a minimally decent life, if society can afford it, and she will be able to improve her situation above this level according to her actual capacities.

VI. FINAL CONSIDERATIONS

I have argued that rational individuals in Rawls' original position would choose the stabilized Nash bargaining solution as a principle of distributive justice for the basic structure of society and, if the individuals are reasonable in the sense that they embrace the moral ideals that underlie the design of the original position, then they also have reasons to follow this principle in their real lives, if others do so, too. Before the stabilized Nash bargaining solution can be applied, however, certain problems must be solved that arise, in one form or another, in the application of any principle of distributive justice. In the following, I address some problems of the stabilized Nash bargaining solution.

First, the members of a society must find agreement on criteria that determine an individual's minimum standard of living and on a way to measure it. These criteria may specify certain social and economic goods that are necessary to live a minimally decent life according to the standards of a particular society. The criteria may include more objective factors that are quantifiable and measurable, such as an individual's income and wealth, but they may also encompass more subjective and, thus, likely more controversial factors, such as an

https://doi.org/10.1017/S0953820810000348 Published online by Cambridge University Press

individual's self-esteem and social status.[32] The main criterion for determining the minimum standard of living in order to secure stable long-term cooperation is that the members of society regard the types of goods and amounts provided as sufficient for living a minimally decent life.

Second, the members of a society must consider as justified their own holdings and the holdings of their conflict partners, which together determine the individuals' relative bargaining power. In other words, the status quo must be justified before the stabilized Nash bargaining solution can be applied as a principle of distributive justice, because the stabilized Nash bargaining solution is, in Nozick's terms, a historical patterned principle of distributive justice.[33]

The stabilized Nash bargaining solution is a historical principle of distributive justice because it makes reference to past events to determine whether the current distribution of goods is justified. It is patterned because it singles out a specific feature of individuals, in this case their basic needs, and the individuals' holdings must be proportionate to this feature up to a certain maximum level. By contrast, the difference principle and the average utility principle are (arguably and respectively, non-patterned and patterned) end-state principles of distributive justice. End-state principles consider only one time slice, namely, the current distribution of goods, and determine on the basis of the information that is given in this time slice whether the current distribution is just. As a consequence, end-state principles do not need to consider the normative status of the status quo as a basis for distributive concerns.

The original position *per se* does not rule out the choice of a historical patterned principle of distributive justice, such as the stabilized Nash bargaining solution, because the individuals under the veil of ignorance know that they will generally be endowed with certain (natural and social) holdings once they leave the original position. As such, the individuals must ensure only that these holdings are not acquired unjustly according to the demands of the principles of justice derived in the original position, and, in order to secure stable long-term cooperation, that each group member reaches at least her minimum standard of living independent of her actual holdings in the real world, as demanded by the stabilized Nash bargaining solution.[34]

[32] Rawls' theory of justice faces a similar problem with regard to determining the criteria that define the worst-off group in society. See Rawls, *Justice as Fairness – A Restatement*, pp. 57–61, for example.

[33] See Robert Nozick, *Anarchy, State, and Utopia* (Oxford, 1974), pp. 153–60.

[34] I think that Nozick's claim in *Anarchy, State, and Utopia*, p. 199, that Rawls' original position rules out the choice of any historical principle of distributive justice is partly

https://doi.org/10.1017/S0953820810000348 Published online by Cambridge University Press

The members of our Rawlsian society must agree on a way to treat historical injustices in the original position. Depending on the precise situation of a society, the problem can be solved in two ways. If the stabilized Nash bargaining solution is applied within an ongoing society in which the group members accept the current distribution of holdings that has evolved within an established legal system after they leave the original position, then this problem is less pertinent. If the group members reject the status quo after the veil of ignorance is lifted, then they may decide on a one-off redistribution, or a gradual redistribution over time, in order to correct past injustices according to the demands of the principles of justice derived in the original position before the stabilized Nash bargaining solution is fully applied as a distributive principle. Because the members of our society are assumed to be reasonable in real life, I am optimistic that they will find agreement on the redistributions to be made, even if the historical information that is necessary to correct past injustices is not fully accessible.

Finally, the stabilized Nash bargaining solution is assumed to guide the arrangement of the basic institutions of society. As such, the demands of the stabilized Nash bargaining solution must be translated from the individual level to the institutional level. The demands are twofold. First, each member of society must be allowed to defend her preferences in conflicts over the cooperative surplus at least as much as possible according to her actual capacities, restricted only by the side-constraint of securing stable long-term cooperation.

This demand seems to be best fulfilled by the institution of a free market system, which tends to distribute the gains of cooperation proportional to the market participants' contribution to the production process, as measured in terms of labor, skills, capital investment, etc., and in this sense according to the market participants' relative bargaining power, loosely speaking. Further, markets seem to be, from a historical perspective, the best mechanisms for allocating and

misguided. The individuals in Rawls' original position do not know their *specific* holdings in real life, but they know that they will be endowed with certain holdings. As such, they must decide on the status of those holdings in the original position, although their decisions under the veil of ignorance are not allowed to be guided by moral considerations. The individuals may regard historical entitlements as irrelevant to their considerations in the original position, and consequently choose an end-state principle of distributive justice. Or, they may regard historical entitlements as relevant to their considerations in the original position, and consequently choose a historical principle of distributive justice, as in our case, which relies on the assumption that instrumental rationality combined with the ideal of individual utility maximization requires agents to defend their preferences in conflicts over the cooperative surplus at least as much as possible according to their actual capacities in real life. This requirement is not a moral assumption, but it allows the choice of a historical principle of distributive justice in Rawls' original position.

https://doi.org/10.1017/S0953820810000348 Published online by Cambridge University Press

distributing scarce resources (with some qualifications concerning the provision of public goods and the use of common goods), because most other mechanisms have shown their inability to maintain mutually beneficial stable long-term cooperation.

The second demand of the stabilized Nash bargaining solution is to grant each member of society at least the means that she needs to maintain her life, assuming that society can afford it. This demand is best fulfilled by the institution of a social welfare net that complements the market system and that assists individuals who cannot fully participate in the market system, or who cannot participate in the market system at all. The types of goods that are included in the social minimum, such as housing, health care, and a basic income, for example, and the precise amounts of these types of goods, depend on a society's definition of an individual's minimum standard of living.[35]

If these problems that the stabilized Nash bargaining solution faces in its practical application can be sufficiently resolved, then the stabilized Nash bargaining solution can serve as a principle of distributive justice in a pluralistic society of rational and reasonable individuals who regard Rawls' original position as a legitimate device to derive principles of justice. If the members of society agree on more specific moral ideals than are expressed by Rawls' original position, then they may, of course, find agreement on a more demanding principle of distributive justice in terms of redistribution and equality than the stabilized Nash bargaining solution.[36]

moehler@vt.edu

APPENDIX

Considering the assumptions made in section III, and restricting the discussion to a two-person society, the situation of the individuals in our society, who face a pure distribution problem, is represented in a stylized form in figure 1.

[35] The resulting form of the state, if the stabilized Nash bargaining solution is implemented as a principle of distributive justice, is best described as a minimal welfare state, in contrast to Rawls' defense of a property-owning democracy and liberal socialism (see Rawls, *Justice as Fairness – A Restatement*, pp. 135–8), although further institutional arrangements may have to be made if Rawls' liberty principle and the fair equality of opportunity principle are also implemented.

[36] I am grateful to Richard Bradley, Mark LeBar, Stuart Rachels, Robert Sugden, Alex Voorhoeve, and the audiences of various conferences and seminars at which I presented earlier versions of this article for their helpful discussions and comments.

https://doi.org/10.1017/S0953820810000348 Published online by Cambridge University Press

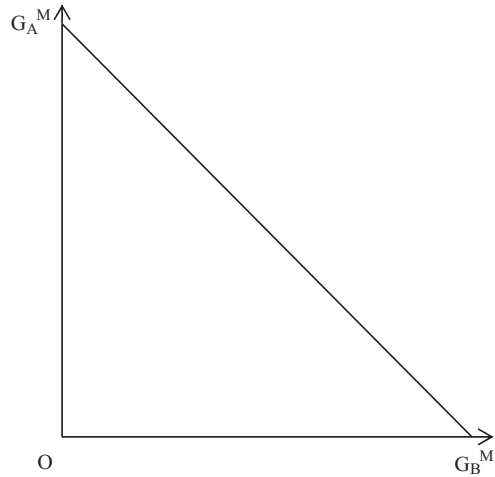


Figure 1.

The axes of the coordinate plane shown in figure 1 represent the amount of the cooperative surplus, expressed in goods, which individual A and individual B can gain if they find a way to resolve their conflict. Increasing amounts of the cooperative surplus are measured upwards from the origin for individual A and rightwards for individual B. If individual A receives the whole cooperative surplus as a result of conflict resolution, then her gain is ${G_A}^M$. Individual B's maximal gain is represented by ${G_B}^M$. All feasible conflict outcomes lie in the area that is depicted by the triangle $O/{G_B}^M/{G_A}^M$. Assuming that each individual aims to gain as much as possible in the process of conflict resolution, all Pareto-optimal outcomes lie on the Pareto frontier, which is the line joining ${G_A}^M$ with ${G_B}^M$.

However, what ultimately counts for rational individuals is not how many units of the cooperative surplus they gain by resolving their conflict, but how much utility they gain by the goods received. The latter depends on the individuals' (von Neumann–Morgenstern) utility functions over the possible conflict outcomes. As such, the goods-space must be linked with the utility-space in the following. To this end, I assume for simplicity that the individuals' utilities increase linearly with any additional unit of the cooperative surplus that they gain, and that both individuals are risk-neutral. Nevertheless, I do not assume that the individuals have identical utility functions, and thus their utilities are not necessarily interpersonally comparable. The individuals' utility functions are assumed to be unique only up to a positive affine transformation.

https://doi.org/10.1017/S0953820810000348 Published online by Cambridge University Press

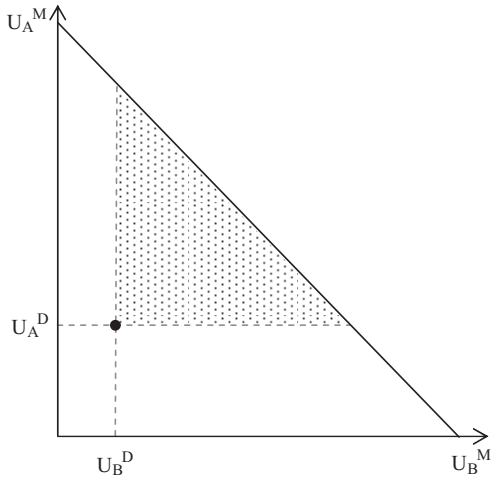


Figure 2.

These simplifying assumptions are not essential to the argument and thus do not restrict the generality of the analysis. They are introduced only to allow a simple graphical representation of the process of conflict resolution. As a result of these assumptions, the graphical representation of the utility-space is identical to the graphical representation of the goods-space, as shown in a stylized form in figure 2.

The axes of the coordinate plane depicted by figure 2 now represent the utilities that the individuals expect to gain if they find a way to share the gains of cooperation. The maximal amount of utility that individual A can gain in the process of conflict resolution is represented by $U_A{}^M$, and for individual B by $U_B{}^M$. The level of utility that an individual must receive in order to cooperate with her opponent in a particular conflict situation is expressed by the disagreement point $U_A{}^D / U_B{}^D$. Assuming that the conflicting parties accept the disagreement point as a basis for conflict resolution, the shaded area of figure 2, which is bounded, convex, and closed, represents the outcomes that allow both parties to improve their situations. If both parties aim to maximize their utility in the process of conflict resolution, then all Pareto-optimal outcomes lie on the Pareto frontier along the shaded area.

In order to resolve their conflict, the individuals must engage in a bargaining process in which their relative bargaining power determines the extent of their gain. Given the simplifying assumptions made by moving from the goods-space to the utility-space, in the described

https://doi.org/10.1017/S0953820810000348 Published online by Cambridge University Press

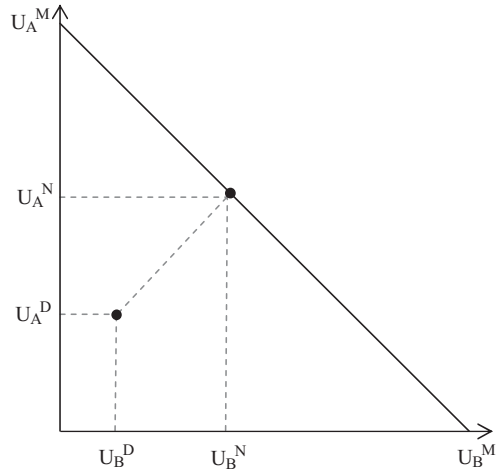


Figure 3.

framework the individuals' relative bargaining power is fully expressed by the location of the disagreement point. In the case depicted by figure 2, this means that individual *A* has more bargaining power than individual *B*, because individual *A* gains more utility, in relative terms, than individual *B* at the disagreement point. Individual *A* can demand more units of the cooperative surplus as a starting point for the bargaining process than her opponent.

As argued, the individuals in our specific situation resolve their conflict according to the Nash bargaining solution. But where does the standard Nash bargaining solution lie graphically within our framework? The standard Nash bargaining solution demands that the individuals maximize the product of their excess utilities above the disagreement point. Although the individuals' utilities are not necessarily interpersonally comparable, their utility scales relative to the disagreement point can be normalized without loss of generality, given the assumption of scale invariance. That is, the individuals' disagreement payoffs can be set to 0, and their utility gains in the bargaining process can be transformed into values between 0 and 1. Assuming this normalized set-up, the product of the individuals' excess utilities above the disagreement point is maximal if both individuals gain an equal amount of utility from the bargaining process. Graphically, the Nash bargaining solution is determined, in the normalized set-up, by the intersection of the 45° angle that passes through the disagreement point with the Pareto frontier, as shown in figure 3.

https://doi.org/10.1017/S0953820810000348 Published online by Cambridge University Press

To clarify, although the individuals gain an equal amount of utility from the bargaining process based upon the disagreement point in the normalized set-up, they do not end up with an equal amount of the cooperative surplus overall. Instead, the stronger individual must secure, in absolute terms, more of the cooperative surplus in order to benefit, in relative terms, as much as her weaker opponent. In our framework, this consideration is expressed by the location of the disagreement point, which allows individual A to demand more of the cooperative surplus at the beginning of the bargaining process than individual B. Expressed in terms of utility, individual A's post-bargaining utility is represented by ${U_A}^N$, and individual B's post-bargaining utility by ${U_B}^N$.

In the following, I illustrate the distributional effects of settling conflicts according to either the standard Nash bargaining solution or the stabilized Nash bargaining solution. To this end, I discuss three different one-off bargaining situations. However, because the stabilized Nash bargaining solution specifies the minimal restrictions on the behavior of rational individuals that must be fulfilled in the short run in order for stable long-term cooperation to be possible, long-run considerations are considered in the short run, if conflicts are resolved according to the stabilized Nash bargaining solution. In particular, the stabilized Nash bargaining solution ensures that each individual receives, in the short run, at least a share of the gains of cooperation that allows her to maintain her life, if the overall cooperative surplus permits it.

The cases that I discuss clarify the distributional differences that arise in the short run if conflicts are resolved according to either the standard Nash bargaining solution or the stabilized Nash bargaining solution. The cases differ in terms of the distribution of power between the conflicting parties, as expressed by the location of the disagreement point, although individual A is always the stronger individual.

Case 1: Assume that individual A is stronger than individual B, but she is neither strong enough to force B to cooperate, nor can she take the life of her opponent. Further, let us assume that the disagreement point in the particular conflict is above each individual's minimum standard of living m, and that the individuals' minimum standards of living, expressed in utilities, are identical (${U_A}^m = {U_B}^m$). Both individuals are at the beginning of the bargaining process at a level that allows them to satisfy their basic needs.

In this case, the standard Nash bargaining solution fulfills the two conditions set out, and therefore the standard Nash bargaining solution N is identical to the stabilized Nash bargaining solution S. In more general terms, as long as the disagreement point lies in the shaded area

https://doi.org/10.1017/S0953820810000348 Published online by Cambridge University Press

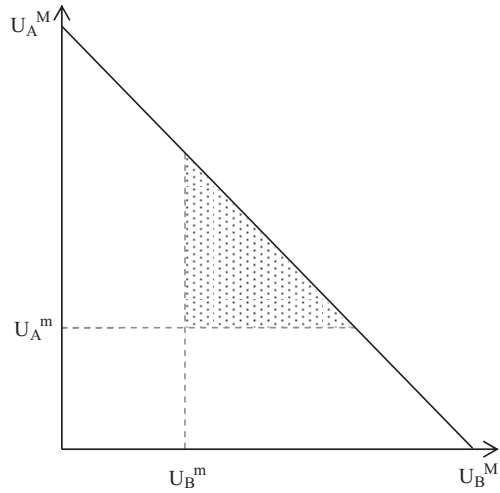


Figure 4.

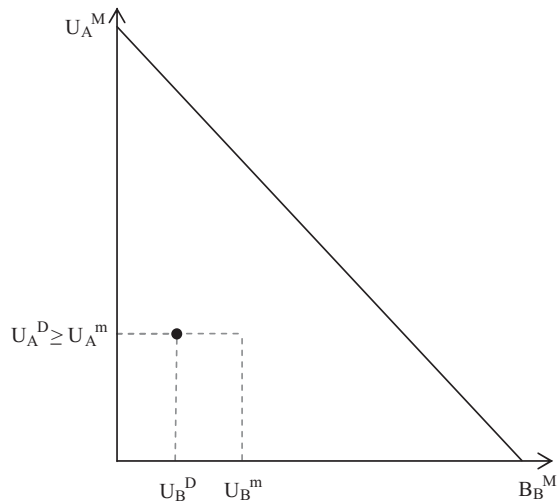


Figure 5.

of figure 4, the standard Nash bargaining solution and the stabilized Nash bargaining solution yield the same bargaining outcome.

Case 2: Now assume that the disagreement point is above, or at least equal to, individual *A*'s minimum standard of living when the conflict arises ($U_A{}^D \geq U_A{}^m$), but it is below individual *B*'s subsistence level ($U_B{}^D < U_B{}^m$), as shown in figure 5.

https://doi.org/10.1017/S0953820810000348 Published online by Cambridge University Press

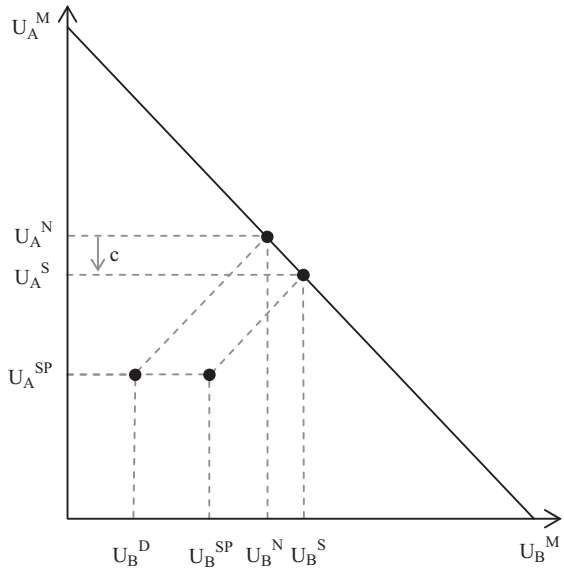


Figure 6.

Let us first address the situation in which the conflict is resolved by the stabilized Nash bargaining solution. According to the stabilized Nash bargaining solution, individual *B* can demand the amount of the cooperative surplus that is necessary to reach her minimum standard of living as a starting point for the bargaining process, if the overall cooperative surplus permits it, because otherwise individual *B* does not agree with the bargaining principle as a principle of distributive justice, as clarified in section II. As such, individual *A* must sacrifice part of her potential gain in order to allow individual *B* to reach her subsistence level as a basis for the bargaining process. The bargaining process occurs not from the disagreement point, but from the stabilized starting point U^{SP}.[37] The latter consists of individual *A*'s disagreement point, because it is above or, as in our case, at least equal to the individual's minimum standard of living ($U_A{}^{SP} = U_A{}^m$) and individual *B*'s minimum standard of living ($U_B{}^{SP} = U_B{}^m$), as shown in figure 6.[38]

[37] The stabilized starting point for the bargaining process reflects the utility levels that the individuals must at least reach in the short run in order to have an interest in stable long-term cooperation, which are the utility levels that reflect the individuals' minimum standards of living.

[38] If both individuals are below their subsistence levels when the conflict arises, then both parties must reach their minimum standards of living before the standard Nash bargaining solution is applied, assuming that the overall cooperative surplus permits it. If both individuals are below their subsistence levels when the conflict arises, and the

https://doi.org/10.1017/S0953820810000348 Published online by Cambridge University Press

In this case, if the conflict is resolved according to the stabilized Nash bargaining solution, the individuals agree on the bargaining outcome $U_A{}^S / U_B{}^S$. If, by contrast, the conflict is resolved according to the standard Nash bargaining solution, the individuals end up with the bargaining outcome $U_A{}^N / U_B{}^N$.

In sum, if the stabilized Nash bargaining solution is applied as a principle of distributive justice, individual *A* must make the additional concession *c* in the process of conflict resolution, which reduces her gain from $U_A{}^N$ to $U_A{}^S$. The size of the additional concession depends on the extent to which individual *B* falls below her minimum standard of living at the beginning of the bargaining process. This brings us to the third bargaining situation.

Case 3: Assume now that the disagreement point is far below individual *B*'s minimum standard of living, and individual *A* is not only strong enough to enforce cooperation with individual *B*, but she can also take the life of her opponent. Individual *A* is a dictator, and thus can take it all. Moreover, let us assume that individual *A* cannot benefit from cooperation with individual *B* in the particular situation. Individual *B* cannot offer individual *A* anything valuable in the short run. Then, if we restrict our analysis to the short run, individual *B*'s relative bargaining power is zero. This third case differs significantly from the two previous cases because, strictly speaking, it represents not a bargaining situation at all because individual *A* does not need individual *B* to cooperate.

The implications of this third case are clear. If the individuals resolve their conflict according to the standard Nash bargaining solution, individual *A* gains the whole cooperative surplus and the bargaining outcome is represented by $U_A{}^M$. In addition, individual *A* can take all of individual *B*'s other goods, including her life, if this action increases individual *A*'s utility. In this case, instrumental rationality requires individual *A* to seize the opportunity, because only then does she maximize her individual utility.

By contrast, if the conflict is resolved according to the stabilized Nash bargaining solution, individual *A* is not able to exert all of her strength on individual *B*. Instead, she must support her opponent so that her opponent can enter the bargaining process from her minimum standard of living. Because individual *A*'s concession, and thus her loss compared to resolving the conflict according to the standard Nash bargaining solution, is highest in case 3, individual *B*'s comparative gain is maximal in this case.

overall cooperative surplus cannot bring both parties to their subsistence levels, then the conflict will be resolved according to the standard Nash bargaining solution.

https://doi.org/10.1017/S0953820810000348 Published online by Cambridge University Press