Philos Stud
DOI 10.1007/s11098-013-0121-y

Uniqueness and symmetry in bargaining theories of justice

John Thrasher

Abstract For contractarians, justice is the result of a rational bargain. The goal is to show that the rules of justice are consistent with rationality. The two most important bargaining theories of justice are David Gauthier's and those that use the Nash's bargaining solution. I argue that both of these approaches are fatally undermined by their reliance on a symmetry condition. Symmetry is a substantive constraint, not an implication of rationality. I argue that using symmetry to generate uniqueness undermines the goal of bargaining theories of justice.

Keywords David Gauthier · John Nash · John Harsanyi · Thomas Schelling · Bargaining · Symmetry

Throughout the last century and into this one, many philosophers modeled justice as a bargaining problem between rational agents. Even those who did not explicitly use a bargaining problem as their model, most notably Rawls, incorporated many of the concepts and techniques from bargaining theories into their understanding of what a theory of justice should look like. This allowed them to use the powerful tools of game theory to justify their various theories of distributive justice. The debates between partisans of different theories of distributive justice has tended to be over the respective benefits of each particular bargaining solution and whether or not the solution to the bargaining problem matches our pre-theoretical intuitions about justice. There is, however, a more serious problem that has effectively been ignored since economists originally

J. Thrasher (✉)
213 Social Sciences, University of Arizona, 1145 E. South Campus Drive,
P. O. Box 210027, Tucson, AZ 85721-0027, USA
e-mail: jthrashe@email.arizona.edu

Published online: 31 March 2013

discussed it in the 1960s, namely the status and implications of the symmetry assumption in all major bargaining solutions.

I will argue that symmetry is a substantive normative constraint that is added into the bargaining procedure, not an implication of standard accounts of rational choice. Introducing such a substantive constraint into the bargaining problem effectively begs the question in favor of some solutions—assuming at the outset what these bargaining theories are attempting to prove. I show that the problems associated with symmetry are present not only in David Gauthier's well-known solution to the bargaining problem, but also in John Nash's solution to the bargaining problem.

1 Bargaining and justice

All bargaining solutions seek a unique solution to the problem of how to rationally divide a surplus of goods or value. To understand the importance of a unique solution for the bargaining problem, it is helpful to look at how bargaining solutions have been used to justify various theories of distributive justice. These justifications, I will argue, do not succeed. They attempt to show that rational agents would uniquely chose one way to divide up a surplus of goods—that only one solution to a problem of division can be proven to be rational. To get a unique solution, however, these theories require the introduction of a symmetry requirement. This symmetry requirement, however, is not an implication of rationality, but is rather a constraint on rational bargaining. As such, it can only be the output of a rational bargain, not an input. To do otherwise would be to assume at the outset what theorists of justice who use a bargaining model are trying to prove, that rational individuals would agree to a specific distributive scheme on purely rational grounds.

Bargaining theories of justice require a unique solution to the bargaining problem, they require that there is one and only one rationally correct conclusion about how to divide the benefits and burdens of social life. I call this property *uniqueness*. Without a unique solution to this distributive question, it will be impossible to justify a particular distributive pattern as the just pattern of distribution. Instead, it will be one among many possible just distributions. Those who are not as well off under distribution A can legitimately argue that distribution B, from their point of view (assuming both A and B are possible solutions to the bargaining problem), is a more just or justified state of affairs. Imagine that both an egalitarian and a utilitarian solution to the bargaining problem are shown to be rational. Why should utilitarians be comfortable living under an egalitarian regime and vice versa? The goal of all roughly contractarian theories that deploy bargaining solutions is to show that their preferred solution is rationally unique.

Rawls recognized this point and even though he specifically rejected bargaining solutions to the problem of justice arguing, "to each according to his threat advantage is hardly the principle of fairness" (Rawls 1958, p. 177). Even so, Rawls recognized that something like the Pareto principle, which only specifies a range of Pareto optimal solutions and does not specify a uniquely optimal solution, would be deficient for a theory of justice. The bargaining problem generates a range of options on the Pareto frontier of two parties, a solution to this problem is supposed

to show why a particular spot on that frontier is rationally required. Any approach that could not generate a unique solution would be incomplete.

I argue in the next section that the introduction of mixed-strategies solutions into games creates a multiplicity of possible solutions. To generate a unique solution requires introducing various refinements to the traditional solution concepts. The most, seemingly, innocuous is the symmetry assumption used by Nash, Gauthier, and virtually every theory of rational bargaining. Symmetry has serious consequences for the contractarian project of showing that justice can be justified through a process of rational bargaining. To show this, I will look at two of the most prominent bargaining solutions, those proposed by David Gauthier and John Nash respectively.[1]

Gauthier contractarian theory of justice, developed in *Morals by Agreement*, is one of the most sophisticated complete contractarian theories. Around a decade after he published *Morals by Agreement*, Gauthier altered his theory in the face of criticism and adopted the Nash's bargaining solution (1993). Further, many contemporary contractarian theorist use the Nash bargaining solution.[2] Because the Gauthier's bargaining solution and Nash's both use a symmetry assumption, I will first look at the Gauthier solution and then turn to the Nash solution. Both are undermined by their shared use of symmetry. Before turning to each bargaining solution, however, it is important to clearly define what symmetry is.

2 Symmetry

In "The Final Problem" Sherlock Holmes is confronted with a strategic dilemma. Professor Moriarty, his arch-nemesis, is attempting to find and kill him. In the past, Holmes easily outsmarted his opponents, not so with Moriarty. Holmes is notoriously vain and unwilling to pile accolades on others, especially for intelligence. Of Moriarty, however, he says:

> He is the Napoleon of crime, Watson. He is the organizer of half that is evil and of nearly all that is undetected in this great city. He is a genius, a philosopher, an abstract thinker. He has a brain of the first order. He sits motionless, like a spider in the centre of its web, but that web has a thousand radiations, and he knows well every quiver of each of them (Doyle 1986 [1893], p. 645).

After being threatened and attacked by Moriarty, Holmes and Watson decide to flee to Europe. They board a train at Victoria Station bound for Dover. Moriarty sees them leaving and tries, in vain, to stop their train. Holmes realizes that Moriarty will rent a special train to overtake them at Dover, concluding that this is what he would

[1] David Gauthier's bargaining solution, minimax relative concession, is a variant of the Kalai-Smorodinsky solution (1975).

[2] Those include, for instance, Ken Binmore (1994) and Ryan Muldoon (2011a, b). Michael Moehler has defended a variant of the Nash bargaining solution, what he calls that stabilized Nash Bargaining solution (2010). H. Peyton Young also defends the Nash bargaining solution as being the most equitable bargaining solution (1995, p. 129).

do in a similar situation. Moriarty is just as smart and knowledgeable as Holmes so he can expect Moriarty to behave the same. This conclusion leads Holmes and Watson to get off the train before Dover at Canterbury to evade Moriarty's special train.

There is something incoherent in Homes's decision to get off at Canterbury. If Moriarty is really the equal of Holmes, why wouldn't he expect Sherlock to evade him by getting off the train early in Canterbury? This problem also puzzled one of the pioneers of game theory, Oskar Morgenstern.[3] Moriarty's decision is based on what he thinks Holmes will do. Holmes's decision is similarly related to what he thinks Moriarty will do. In considering the Holmes problem in his 1928 book, Morgenstern came to a striking conclusion:

> I showed in some detail in particular that the pursuit developing between these two [Moriarty and Holmes] could never be resolved on the basis of one of them out-thinking the other ("I think he thinks that I think! !…"), but that a resolution could only be achieved by an "arbitrary decision," and that it was a *problem of strategy* (1976, p. 806 emphasis added).

John von Neumann took up this "problem of strategy" with Morgenstern and together they developed the basis of what would become game theory. In *The Theory of Games and Economic Behavior*, they modeled the Holmes-Moriarty problem as a zero-sum game (2007, pp. 176–178).[4]

The complexity of the example arises from the symmetry of the parties involved. Holmes cannot out-think Moriarty because Moriarty is effectively his rational twin. The solution to this problem is to break the symmetry by introducing randomness. If not even Holmes knows what he is going to do, Moriarty will not either. This solution was developed into the idea of a mixed-strategy. In some games, it is beneficial for individuals to choose a strategy randomly. Randomness breaks the symmetry and makes one's action harder to predict. Mixed-strategies also have the effect of introducing many potential strategies where there were none before. As such, it makes the search for a single optimal equilibrium solution more difficult.

The pioneers of bargaining theory saw this as a problem. Their goal was to find one and only one rational solution to what they called "the bargaining problem." This is the problem of deciding how to divide a set of goods where no party has any antecedent claim and where any mutually agreed upon decision will be binding. Nash uses the example of a labor union negotiating with a firm as an example of

[3] For a discussion of the importance of this earlier work to his later collaborative work with John von Neumann see: (Morgenstern 1976, p. 806; Innocenti 1995).

[4] Specifically, they modeled it as a form of "matching pennies." The normal form of the game is below:

		Holmes	
		Canterbury	Dover
Moriarty	Canterbury	(100, -100)	(-50, 50)
	Dover	(0,0)	(100, -100)

such a problem (1950, p. 155). Nash's goal was to articulate a unique "solution" to this problem. He explains:

> It is the purpose of this paper ["The Bargaining Problem"] to give a theoretical discussion of this problem and to obtain a definite "solution" making, of course, certain idealizations in order to do so. A "solution" here means a determination of the amount of satisfaction each individual should expect to get from the situation, or, rather, a determination of how much it should be worth to each of these individuals to have this opportunity to bargain (Nash 1950, p. 155).

Nash proved that his solution to the bargaining problem uniquely satisfied four simple axioms (1950). John Harsanyi later extended Nash's solution (1956, 1958). One of Nash's axioms is "symmetry." This assumption is very similar to the assumption that motivated the Holmes-Moriarty problem, namely that both parties are equally rational and well informed. Symmetry rules out any asymmetrical solutions to the bargaining problem.

The symmetry axiom is defined progressively over the course of several of Nash's early articles. In his first paper on the bargaining problem, Nash defines symmetry laconically as expressing "equality of bargaining skill" (1950, p. 159). Nash clarifies and, importantly, changes this definition in his 1953 paper, writing:

> The symmetry axiom, Axiom IV, says that the only significant (in determining the value of the game) differences between the players are those which are included in the mathematical description of the game, which includes their different sets of strategies and utility functions. One may think of Axiom IV as requiring the players to be intelligent and rational beings. But we think it is a mistake to regard this as expressing "equal bargaining ability" of the players, in spite of a statement to this effect in "The Bargaining Problem." With people who are sufficiently intelligent and rational there should not be any question of "bargaining ability," a term which suggests something like skill in duping the other fellow. The usual haggling process is based on imperfect information, the hagglers trying to propagandize each other into misconceptions of the utilities involved. *Our assumption of complete information makes such an attempt meaningless* (Nash 1953, pp. 137–138 emphasis added).

Introducing randomness in the form of mixed-strategies breaks symmetry in non-cooperative games like the Holmes-Moriarty game and introduces a multiplicity of solutions. In bargaining problems theorists have the opposite problem; there are too many potential solutions. In order to find a unique solution, all but one needs to be ruled out. Symmetry helps do this. In effect, it turns every bargaining problem into the cooperative equivalent of the Holmes-Moriarty problem. The bargaining problem differs in that it is mutually beneficial, not zero-sum. We can think of the bargaining problem as Holmes bargaining with his brother Mycroft. Both are equally intelligent and equally knowledgeable. Since both are symmetric reasoners, solutions should also be symmetric. This simplifies the choice problem considerably. Once we know what the endowments are (specifying the threat point) and the

surplus that the agents are bargaining over, we can show what it would be rational to agree to in the bargain.

So, in (some) non-cooperative games symmetry is the problem, whereas in the bargaining problem it is the solution to a problem. In some non-cooperative games with symmetry we often have no possible solutions so we introduce randomness to break symmetry. In the bargaining problem, however, any division of the goods that leaves each party better off than their initial threat-point is beneficial and, hence, a possible solution. The problem here is too many solutions. We can reintroduce a symmetry assumption to narrow down the range of possible solutions to one by modeling the bargaining parties as rational twins.[5] The introduction of symmetry in bargaining solutions is not a minor thing; it is essential to generating a unique solution. In the next two sections, I will look at two particular uses of symmetry in contrarian theories of justice. First Gauthier's use of the assumption in *Morals by Agreement* and then the Nash bargaining solution used by many other contemporary contractarian theorists.

3 Gauthier and symmetry

Symmetry is introduced formally as a condition of Gauthier's adaptation of the Kalai-Smorodinsky bargaining solution: minimax relative concession (Kalai and Smorodinsky 1975, pp. 513–518; Gauthier 1986, pp. 113–156). He describes symmetry as an "equal rationality" condition. Gauthier uses a bargaining model to represent his solution, but as Nash and Ariel Rubinstein have shown, the bargaining problem can also be represented as a non-cooperative game (Rubinstein 1982).[6] A simple bargaining problem can be represented as an asymmetric coordination game where there is no unique solution, such as in the meeting game represented below in Fig. 1. Both parties have reason to coordinate on the same solution. There is, however, disagreement about which solution is preferable. Consider the simple meeting game in Fig. 1. In this game, both parties prefer to meet but have different preferences over where they would like to meet. Their preferences are indicated by Roman numerals where I > II > III. Row, a lover of interesting and exotic beers, would rather meet at the bar. Column, who loves the outdoors, would rather meet in the park. Deciding where to meet in this situation will involve one party making a concession to the other.

If coordination is over issues concerning justice, for instance the choice of property regimes, the situation is similar. Each party may prefer some property arrangement rather than none at all.[7] Each party, however, has a preferred property

[5] Binmore also uses the language of twins in his discussion of the "paradox of the twins" and the "symmetry fallacy" (1994, pp. 203–256).

[6] In general, Rubinstein has shown that the Nash bargaining problem can be represented as a non-cooperative game. The basic idea that I am taking from Rubinstein is that solutions to bargaining problems can often be represented as equilibrium selection problems in non-cooperative games. Nash makes a different point in his 1953 article, but it is instructive that he also models one version of the bargaining problem as a multi-stage threat game.

[7] For Gauthier, the benefits of some system of constraint arise because of the probability of "market failures" in the use of individual reason that lead to prisoner's dilemma like situations (1986, pp. 84–85).

Fig. 1 Ordinal meeting game

	Bar	Park
Bar	I, II	III, III
Park	III, III	II, I

system. As in the meeting game, concessions are required. Consider the example below where two agents are trying to coordinate on the basic institutions of their economic system, in this case two different systems of property ownership (Fig. 2).

Here, the object of coordination is much more substantial than in the previous case. Row and Column are deciding over what rules for property holdings they should have. They must agree, in this game, to have any particular system of property. This excludes the non-coordination outcomes represented in the southwest and northeast quadrants. Row prefers property system A and Column prefers property system B. Given that someone will have to make a concession to generate agreement, how much concession is rational? Any theory of rational bargaining must give a unique answer to this question

Most solutions, Gauthier's included, rely on a mixed-strategy solution to the bargaining problem. To introduce mixed-strategy solutions, we will need to represent the property game with cardinal utilities as in Fig. 3 below.

Each player's benefit is a representation of how they rank the respective outcomes, without the need for interpersonal comparisons. In the mixed strategy solution, each player gets their most preferred option 2/3 of the time and their least preferred option1/3 of the time. The payoffs for each player, as a representation of their preference orderings, are below in Fig. 4.

The mixed-strategy solution breaks the symmetry of the two pure strategies by introducing a suboptimal mixed-strategy solution. This solution breaks symmetry by introducing randomness but it is also "symmetric" in another sense because each player receives the same payoff. It doesn't matter whether either player is Row or Column, each will receive the same payoff in the mixed-strategy symmetric solution. Both players are worse off in the new symmetric solution, however, than they would be in any particular pure strategy solution.

Fig. 2 Ordinal property game

	Property A	Property B
Property A	I, II	III, III
Property B	III, III	II, I

Fig. 3 Cardinal property game

	Property A	Property B
Property A	2, 1	0, 0
Property B	0, 0	1, 2

	{A, A}	{B, B}	Mixed Strategy
Row	2	1	2/3
Column	1	2	2/3

Fig. 4 Cardinal property game expected payoff table

Two problems arise, however, when we try to apply mixed-strategy solutions to questions of justice as in the property game above. How do we make sense of randomizing—either psychologically or in terms of rationally justifying a particular solution in the bargaining problem? In the Holmes-Moriarty game, the rationale for introducing randomizing is strategic. Holmes randomizes to keep Moriarty off balance. There is no analogue to this in the bargaining problem since coordination and agreement is the goal. This undermines the justificatory power of mixed-strategy solution to the bargaining problem. In addition, any proposed solution must also make sense; the contractors must be able to follow the line of reasoning that led to the solution (Pettit 1996, pp. 295–296; Pettit and Sugden 1989). Unlike in the zero-sum Holmes-Moriarty game, where the point is to evade one's opponent, in the coordination game the point is to coordinate on the same solution. What could possibly be the reasoning that would lead them to conclude that they should randomize their behavior so as to chose property system A some of the time and property system B the other percentage of the time, recognizing that they will miss each other altogether a non-negligible amount of the time. The real absurdity of this comes out clearly in Fig. 4. Both would be better off agreeing to either of the pure solutions rather than agreeing to the mixed-strategy.

The second problem is that Nash only proved that there is *at least* one solution to non-cooperative games—often there are many. In some games, there are a huge number of solutions.[8] For contractors to have a reason to make a particular concession, the theorist needs to show that there is a uniquely rational solution where each contractor is making exactly the appropriate concession and no more. Much of bargaining theory is driven by the need to show that a particular solution is uniquely rational; 20th century political philosophers continued this project, sometimes unknowingly. Gauthier, for instance, claims that his solution—minimax relative concession—is uniquely rational (Gauthier 1986, p. 139). Ken Binmore disagrees and thinks a version of the Nash bargaining solution is uniquely rational (2005). Rawls had a different solution.[9] Each theory has its own *uniquely rational* solution. There are many ways to solve the bargaining problem; Gauthier's is one

[8] Consider the ultimatum game or the Nash demand game where every matching solution is a Nash equilibrium.

[9] Of course, Rawls claimed that the difference principle is not the result of bargaining in the traditional sense. This is partly because choice from behind the veil of ignorance is the choice of one person. He continues to use the language of "parties" left over from earlier formulations, however. In Sect. 3 of *Theory*, he claims "the principles of justice are the result of a fair agreement or bargain" (Rawls 1999, p. 11).

Fig. 5 Joint mixed strategy expected payoff table

	{A, A}	{B, B}	Joint Mixed
Row	2	1	1 ½
Column	1	2	1 ½

among many. It is not even the most popular. That honor goes to the Nash solution, which even Gauthier later adopted (1993). Without a unique solution, no party has a reason to prefer one and only one solution. Even if a solution were reached, it would not be rationally defensible and would lack normative force. Unless the solution is rationally defensible, it will not be clear why these and not some other concession are justifiable. It would not show "you and me" why we have reason to endorse and adopt a disposition to be constrained by the rules agreed upon by the contractors. Or, put differently, if we imagine the bargainers as our representatives, even if they reach an agreement we will not have reason to endorse or ratify their agreement. This is why uniqueness is so important and why all attempts to generate rational determinacy in agreement seek a unique solution. The traditional way to solve the uniqueness problem is to introduce refinements to the model of rationality to help choose between multiple bargaining solutions or equilibria.

In *Morals by Agreement,* Gauthier introduces what he calls a joint mixed strategy as a possible way of solving this problem (Gauthier 1986, p. 120). Each party agrees to take their most preferred solution 1/2 of the time. The payoff table for this approach is in Fig. 5.

The basic idea is to turn the simple property game into a two-stage game.[10] In the first stage, the parties agree to use of public mechanism for generating correlation. In Gauthier's case, this could be the public flip of a coin. Both could agree on Property system A if the coin lands on heads and Property system B if it lands on tails. In the second stage of the game, after the coin is flipped, each chooses the property system based on what the correlating mechanism (coin flip) indicates. If they do this, they will get their preferred outcome half of the time and their less preferred outcome the other half of the time. More importantly, they will avoid the missing each other (the non-coordination outcome) altogether, unlike in the mixed strategy solution. This makes the joint strategy preferable to the symmetric mixed strategy. The problem, however, is that the joint solution is a combination of two games: a game to decide on the conditions of correlation and the original property game (Binmore 1993, pp. 137–138). The solution to the first game is just as problematic as the solution to the second game. To solve the first game, Gauthier must also rely on a symmetry condition.

As we can see, symmetry becomes essential to solving this bargaining problem. Both Harsanyi and Rawls—in different ways—also concede that a symmetry

[10] This solution is similar in some ways to the correlated equilibrium solution that Herbert Gintis, following Robert Aumann, has proposed to solve similar indeterminacy problems in another context. Gintis introduces the solution concept to solve certain problems that arise when common knowledge does not obtain and it is appropriate in his context, but is inappropriate here (2009, 2010).

assumption is necessary.[11] For Rawls and Harsanyi this assumption is less problematic because they are explicitly modeling fair or reasonable agreement. Symmetrical solutions will seem fairer because they do not privilege one party over another.

Of course, for those concerned with the contractarian justificatory problem like Gauthier, the fact that symmetrical solutions are fairer does not justify symmetry. To do so would only beg the original question of what system of justice rational individuals would agree to. Gauthier admits as much when he writes, "were I to become convinced that an appeal to equal rationality [symmetry] was either a concealed moral appeal, or inadmissible on some other grounds, then I should have to abandon much of the core argument of *Morals by Agreement*" (Gauthier 1986, p. 186).[12] Gauthier later saw that his particular "equal rationality" or symmetry assumption really was unjustified and he rejected it along with his bargaining solution in favor of the Nash solution (1993, p. 180). Gauthier and many contemporary contract theorists believe that the Nash solution does not have a similarly unjustified assumption, that symmetry in Nash is somehow different from symmetry in Gauthier. As I will argue in the next section, this is unwarranted. Symmetry in Nash is just as problematic as it is in Gauthier's theory, with the same effect.

4 Nash and symmetry

Contemporary contract theorists contend that the Nash solution is immune from the problems that plagued Gauthier. For instance, Ken Binmore, Michael Moehler, Ryan Muldoon, and H. Peyton Young all argue that the Nash bargaining solution or some related variant is appropriate for modeling justice (Binmore 1994; Binmore 2005; Moehler 2010; Muldoon 2011a, b; Young 1995, Chap. 7). While, for Gauthier, symmetry is clearly a moral constraint representing something like fairness or impartiality and is hence in conflict with his project, these thinkers believe that in the Nash solution symmetry is not a moral premise but an implication of rationality. This is complicated since Gauthier's bargaining solution is a variant of the Kalai-Smorodinsky bargaining solution that explicitly uses Nash's symmetry axiom. If Gauthier's theory is susceptible to problems based on symmetry, Nash's should be as well. The difference might be with Gauthier's justification of symmetry in terms of an equal rationality assumption, which he admits is moralized and goes beyond the implications of rationality (1993, p. 180). Many seem to assume, however, that there is some fundamental difference between Gauthier's symmetry assumption and Nash's. There is no warrant for this assumption, as I will show in the rest of this section. Any problem with Gauthier's bargaining theory that arises from the symmetry assumption should apply equally to a theory that uses the Nash solution.

[11] Harsanyi is explicit about this (1982, p. 49) and Rawls makes it clear he is relying on symmetry in *Political Liberalism* (1996, p. 106).

[12] I thank an anonymous referee for alerting me to this point.

The Nash solution's symmetry assumption, according to this interpretation, is not an assumption of "equal bargaining power" as it is described in Nash's early article, but rather a rational condition of any bargaining solution that the solution not vary based on the names or the labels of the bargainers involved. The implication is that the symmetry condition in Nash's solution is fundamentally different from Gauthier's equal rationality assumption. This claim is mistaken. Nash's solution, as many early commenters noted, does employ a symmetry assumption that goes far beyond any straightforward understanding of rationality, something that John Harsanyi clearly understood and defended (Harsanyi 1961; Harsanyi 1982). It is not a demand of rationality, but is a substantive *constraint* on rationality. As such, it should properly be part of the output of a rational contractarian bargain, not one of the inputs.

Recall Nash's expansion of his definition of symmetry quoted above. Nash explains that Axiom IV (symmetry) postulates both perfect information and rationality between the bargainers. This excludes the "usual haggling process" involved in typical negotiations, which Nash describes as "meaningless" (1953, p. 138). As Harsanyi developed the idea, this assumption has the effect of restricting the variables that are taken into account in the bargaining decision rule. Harsanyi writes:

> As any theory must apply to both players, if the two players happen to be equal with respect to all relevant independent variables they must be assigned full equality also with respect to the dependent variables, i.e., with respect to the outcome. But this is precisely what the symmetry postulate says. Different theories of bargaining may differ in what variables they regard as the relevant independent variables but, if the two players are equal on all variables regarded by the theory as relevant, the theory must allot both players the same payoffs (Harsanyi 1961, p. 189).

The purpose of restricting the relevant variables to the bargaining solution is to generate a unique result. As he writes in the same paper, "the symmetry postulate has to be satisfied, as a matter of sheer logical necessity, by any theory whatever that assigns *a unique outcome* to the bargaining process" (Harsanyi 1961, p. 188 emphasis added). We can agree with Harsanyi that symmetry is necessary for generating a unique solution without thinking that the symmetry assumption is thereby justified or logically necessary. The question here is whether symmetry is a natural implication of rationality or whether it is an antecedent constraint on rationality.

5 Against symmetry

Thomas Schelling argued that the symmetry assumption is not a condition of rationality but rather a constraint that is often not justified and is certainly not logically necessary (1959, p. 219). It is not a direct implication of rationality, nor is it justified on strategic grounds that there are always benefits to reasoning symmetrically. It is rather an artifact of the mathematical obsession with

uniqueness. It is one of the supposed benefits of the Nash solution that it can be translated into non-cooperative game theory. This is meant to show that the solution does not have moral content but is instead the result of strategic rationality consistently applied. Schelling disagrees and argues:

> it is not a universal advantage in situations of conflict to be inalienably and manifestly rational in decision and motivation. Many of the attributes of rationality…are strategic disabilities in certain conflict situations. It may be perfectly rational to wish oneself not altogether rational, or—if the language is philosophically objectionable—to wish for the power to suspend certain rational capabilities in certain circumstances (Schelling 1960, p. 18).

Indeed, Schelling goes on to gives two examples where rationally reducing one's rationality could be strategically beneficial. The first is in cases of intimidation or extortion. Caesar burned his ships on the beach after reaching Britain in order to credibly threaten the Britons and to indicate to his men that the only strategy is to fight with no hope of retreat. In the second case, negotiation skill can benefit from reducing one's options. As Schelling writes, "the power of the negotiator often rests on a manifest inability to make concessions and demands" (1960, p. 19). Negotiations, as a game of threats, are often games of chicken. In those games, it can be beneficial to limit one's options in order to credibly threaten the other party. That is, it is possible and often beneficial to strategically reduce one's rational options asymmetrically.

Consider a similar case used by Derek Parfit (1987, 12–13). In this case, a man breaks into my house and orders me to open my safe in order to steal the gold I have stored there. He knows the police will not arrive in time to stop him and that I will not give up the gold voluntarily. He threatens to shoot my children one by one and to torture me until I open the safe. I know that if I give him the money, he will kill all of us to eliminate any witnesses. What am I to do? Parfit argues, following Schelling, that if there were a drug that I could take to make myself irrational, I would be justified in taking it. This would make me act manifestly irrational telling the burglar "Go ahead. I love my children. So please kill them" (Parfit 1987, p. 13). The burglar seeing that I am irrational, would know that there is no hope of getting the gold out of me and presumably leave. Parfit concludes "it would be rational for me, in this case, to cause myself to become for a period irrational" (ibid.).

In all of these cases there are compelling rational strategic reasons to behave irrationally or to commit to an irrational strategy. This is especially relevant in real negotiations and bargains. Richard Nixon, for instance, employed what he called a "madman theory" to bring the North Vietnamese to the bargaining table by convincing them and the Soviets that he was unstable and irrationally afraid of communism. If he thought he was losing the war in Vietnam, he would be prepared to use nuclear weapons to win the conflict. He believed this would bring the North Vietnamese to the bargaining table. Symmetry is clearly neither a "logically necessary" condition of rationality, nor is it justified on strategic grounds. One is not necessarily being inconsistent or imprudent by reasoning asymmetrically.

These considerations led theorists like Schelling away from a search for rationally unique solutions to games and bargaining problems and towards a

psychological and empirical study of focal points. As Schelling argues, the Nash solution, with its symmetry assumption "is limited to the universe of mathematics…which should not be equated with the universe of game theory" (Schelling 1960, p. 290). It is true that "mathematical esthetics" requires uniqueness in it solutions, but this fact does not guarantee that uniqueness is a property of game or bargaining solutions (ibid.). The existence of a unique solution must be a conclusion and should not be assumed as a *premise*. An assumption of symmetry cannot be justified on the basis that it generates unique results unless we have some antecedent reason for thinking that unique solutions should always be forthcoming.

Schelling criticisms form two arguments against symmetry, each of which should lead to a rejection of the axiom on rational grounds. First, symmetry is not a condition of rationality, but rather a constraint on rationality. It is not "logically necessary" for rational agents to reason symmetrically. For this reason we should reject the symmetry axiom. Second, the only reason we have to endorse symmetry is to generate a unique solution to the bargaining problem. Uniqueness, as Schelling argues, is something we prove or discover, not something we assume is always possible. This justification for symmetry also goes beyond the assumptions of rationality and cannot be endorsed by the contractarian.

Schelling's point is even more important if bargaining problems and games are meant to model agreement over the rules of justice. Uniqueness may be mathematically important, but there is no reason to think that justice is something that has a unique solution. These bargaining models are mathematically interesting but, as Ariel Rubinstein would no doubt be the first to point out, have very little to do with the process of actual bargaining or negotiation. This is relevant because the point of modeling contractual agreement over rules of justice as a bargain is to model the relevant rational features of actual persons. Unless we model our contractors as optimizers who are constrained by the symmetry assumption of the Nash bargaining theory, there is no reason to think that it will be strategically beneficial to reason symmetrically or restrict their agreement to symmetrical solutions.

There are two main objections to this rejection of symmetry as an implication of rationality, one based on stability concerns, the other based on the nature of the contractual model. First, one can argue that symmetry may not be a direct implication of rationality, but that it is an indirect condition of rationality since only symmetric bargains will be stable. Symmetry might be justified as a requirement of stability. In the symmetrical solution to the property game, for instance, each party gets his or her way some of the time. In that sense, the mixed strategy solution is fairer. It might be argued that because it is fairer, it will be more stable.[13] If it were more stable, this would give rational maximizers in the agreement situation good reason to prefer symmetrical solutions to non-symmetrical solutions.

There are three reasons why this indirect way to introduce symmetry fails. First, it is important to remember that all solutions to either the bargaining problem or the property game are *equilibria*, that is, they are all by definition stable—at least in the technical sense that no player has reason to unilaterally switch to a different

[13] This is exactly the kind of justification that Binmore makes to defend the evolutionary salience of fair social contracts in *Natural Justice* (Binmore 2005).

strategy. This holds true for the mixed-strategy equilibrium too, even though it is Pareto dominated (everyone is worse off).

Second, why should the parties to the agreement, modeled as rational optimizers, think that the fairer solution is more stable? There is no reason to think that they have a sense of fairness. Further, according to the defenders of the Nash solution, it does not involve any substantive considerations of morality or fairness. For this argument to have teeth, the defender of symmetry must admit what he cannot admit: that symmetry is a condition of fairness.

This raises an important point about the nature of each bargainer's utility function, namely that they need not be purely self-interested, they can take any form. Imagine the case of two pure altruists. Both have preference orderings that favor the other person getting what they want most of the time. For both agents to be satisfied, however, each will need to make a concession on how much they help the other. Additionally, even though each party's preferences are other-regarding, they are still asymmetric in the way that agents in the property game are and, hence, a similar conflict would arise regardless of the content of their actual preferences.

Third, the available experimental evidence is not consistent with the claim that symmetrical solutions are more stable. This was, in effect, Schelling's point. Andrew Schotter and Barry Sopher, for instance, have shown experimentally that in a game formally identical to the property game played over several "generations," where players can communicate and advise one another, non-symmetrical equilibria become the focal points (Schotter and Sopher 2003, p. 513). Apparently, players settle on asymmetric equilibria and then pass those solutions on to the next generation. This suggests that asymmetric equilibria are often stable and passed on through social learning and imitation. It is true that in repeated bargaining and ultimatum games, norms of fairness and symmetry can emerge in certain circumstances, but they are not guaranteed to emerge and it is only rational to make and accept symmetrical bargains if others are doing so (Bicchieri 2006, pp. 222–225). Symmetrical solutions can become norms, but they are chosen because they are norms not because they are symmetrical. Non-symmetrical solutions can also become norms.

An even more serious problem involves common knowledge. For symmetry to be a condition of rationality, assuming that all of the other arguments I have given against it have failed, we must still assume that the symmetry of the agents in the game or bargain is common knowledge. The contractors must reason symmetrically and know that every other contractor reasons symmetrically and know that every other contractor know, etc. This is an incredibly high epistemic burden to bear; there is no reason to think that we should model the contractual agreement situation as an environment with common knowledge of symmetry, unless it was a necessary requirement of strategic rationality. There is, however, nothing in the representation of the parties as rational optimizers that should make us think either that (a) they will reason symmetrically or (b) that they will have reason to believe that all other contractors will reason symmetrically. Even if the party in question reasons symmetrically, this is not necessarily good evidence that others are symmetrical reasoners. The contractor would have to know antecedently that the other parties are symmetrical reasoners. Otherwise, the other parties could pretend to be symmetrical

reasoners and get the other party to make unnecessarily high concessions. This creates a potential for the agreement situation to turn into an epistemic prisoner's dilemma over whether to behave symmetrically. Without assuming that symmetry is a *constraint* on reasoning, reasoning symmetrically would open one up to the possibility of exploitation.

I have argued that symmetry cannot be indirectly justified as a condition of rationality because rational bargainers would endorse it for stability reasons. Further, the assumption also requires an additional assumption that symmetry is common knowledge. Another indirect approach to justify the rationality of symmetry is to argue that the contractual model itself requires symmetry. That to generate normative agreement over rules of justice, a bargain needs to be fair in some relevant sense and that fairness can be represented in a thin way by a symmetry condition. Notice that this justification admits that symmetry is not a direct condition of rationality, but rather is justified by rational contractors who are meant to agree on fair rules of justice. Many early commentators on the Nash bargaining solution seemed to agree with this point, as Alessandro Innocenti points out, arguing that symmetry could "only be accepted as an ethical criterion" (Innocenti 2008, Sect. 4). Martin Shubik, for instance, writes, "the Nash solution to the bargaining problem suggests a method of 'fair division.' The best way to look at the motivation behind this method is that it is normative" (Shubik 1959, p. 49). Luce and Raiffa make a similar point in *Games and Decisions* (1957, pp. 135–137).

Most contemporary thinkers disagree with these early critics, arguing that the Nash solution has "no merit as an ethical concept"(Binmore 1994, p. 83). Binmore notes, rightly, that if the Nash solution did have ethical merit, this would disqualify it from being used by contractarians. To hold this view, though, one must disregard the concerns with the rationality of the symmetry assumptions discussed above. The restrictions on the bargaining situation that are necessary to generate the Nash solution—that bargainers be limited to symmetrical solutions and reason in exactly the same way as their fellows—are excessive and unjustified. Unless we assume that our contractors are somehow obsessed with following the reasoning of Harsanyi, Zeuthen, and Nash there is no reason to think they should abide by a symmetry constraint. As we have already seen, many—most notably Schelling—do not believe that something as strong as the symmetry condition can be generated out of the assumptions of optimizing rationality alone.

H. Peyton Young justifies the Nash solution as the only consistent bargaining solution that is impartial and equitable. He argues that the "Nash standard is the most satisfactory way of defining a fair bargain" (Young 1995, p. 122). This may be correct, but since the Nash solution introduces a condition which cannot be justified as an implication of rationality and can only be considered as an, albeit thin, normative constraint on the rationality of the contractual parties, the Nash solution fares no better than Gauthier's original bargaining solution.

The conclusion to draw from all of this is that the most popular bargaining solutions all assume a substantive normative constraint at the outset. As such, they cannot be used to show that rational individuals would choose a particular unique solution in a bargaining situation. This is especially important for theories of distributive justice that attempt to, as Rawls put it, generate the reasonable out of the

rational. That project, insofar as it uses one of the main bargaining solutions or something similar, is unlikely to be successful.

Acknowledgments Special thanks are due to Jerry Gaus and David Schmidtz for their helpful comments on earlier versions of this paper. I would also like to thank Steve Wall, Uriah Kriegel, David Copp, Chris Morris, Ryan Muldoon, Chris Freiman, Kevin Vallier, Keith Hankins, Danny Shahar, Chad Van Schoelandt, Victor Kumar, Michael Bukoski, Bill Glod, Mark Budolfson, and an anonymous referee for comments on earlier versions of this paper.

References

Bicchieri, C. (2006). *The grammar of society: The nature and dynamics of social norms*. Cambridge: Cambridge University Press.

Binmore, K. (1993). Bargaining and morality. In D. Gauthier & R. Sugden (Eds.), *Rationality, justice and the social contract: Themes from morals by agreement* (pp. 131–156). Ann Arbor: University of Michigan Press.

Binmore, K. (1994). *Game theory and the social contract*: *Playing fair* (Vol. 1). Cambridge: The MIT Press.

Binmore, K. (2005). *Natural Justice*. New York: Oxford University Press.

Doyle, S. A. C. (1893). *The final problem. Sherlock Holmes: The complete novels and stories* (Vol. 1, pp. 642–660). Bantam Classics.

Gauthier, D. (1986). *Morals by agreement*. Oxford: Clarendon Press.

Gauthier, D. (1993). Uniting separate persons. In D. Gauthier & R. Sugden (Eds.), *Rationality, justice and the social contract: Themes from morals by agreement* (pp. 176–192). Ann Arbor: University of Michigan Press.

Gintis, H. (2009). *The bounds of reason: Game theory and the unification of the behavioral sciences*. Princeton: Princeton University Press.

Gintis, H. (2010). Social norms as choreography. *Politics, Philosophy, and Economics, 9*, 251–264.

Harsanyi, J. (1956). Approaches to the bargaining problem before and after the theory of games: A critical discussion of Zeuthen's, Hicks', and Nash's Theories. *Econometrica: Journal of the Econometric Society, 24*, 144–157.

Harsanyi, J. (1958). Notes on the bargaining problem. *Southern Economic Journal, 24*, 471–476.

Harsanyi, J. (1961). On the rationality postulates underlying the theory of cooperative games. *Journal of Conflict Resolution, 5*, 179–196.

Harsanyi, J. (1982). Morality and the theory of rational behavior. In A. Sen & B. Williams (Eds.), *Utilitarianism and beyond* (pp. 39–62). Cambridge: Cambridge University Press.

Innocenti, A. (1995). Oskar Morgenstern and the heterodox potentialities of the application of game theory to economics. *Journal of the History of Economic Thought, 17*, 205–227. doi:10.1017/S1053837200002601.

Innocenti, A. (2008). Linking strategic interaction and bargaining theory: The harsanyi-schelling debate on the axiom of symmetry. *History of Political Economy, 40*, 111–132.

Kalai, E., & Smorodinsky, M. (1975). Other solutions to Nash's bargaining problem. *Econometrica, 43*, 513–518.

Moehler, M. (2010). The (stabilized) Nash bargaining solution as a principle of distributive justice. *Utilitas, 22*, 447–473. doi:10.1017/S0953820810000348.

Morgenstern, O. (1976). The collaboration between Oskar Morgenstern and John von Neumann on the theory of games. *Journal of Economic Literature, 14*, 805–816. doi:10.2307/2722628.

Muldoon, R. (2011a). Justice without agreement. Unpublished Manuscript.

Muldoon, R. (2011b). The view from everywhere. Unpublished Manuscript.

Nash, J. (1950). The bargaining problem. *Econometrica: Journal of the Econometric Society, 18*, 155–162.

Nash, J. (1953). Two-person cooperative games. *Econometrica: Journal of the Econometric Society, 21*, 128–140.

Parfit, D. (1987). *Reasons and persons* (paperback edition). Oxford: Oxford University Press.

Pettit, P. (1996). *The common mind: An essay on psychology, society, and politics*. USA: Oxford University Press.

Pettit, P., & Sugden, R. (1989). The backward induction paradox. *Journal of Philosophy, 86*, 169–182.
Rawls, J. (1958). Justice as fairness. *The Philosophical Review, 67*, 164–194.
Rawls, J. (1996). *Political liberalism* (paperback). New York: Columbia University Press.
Rawls, J. (1999). *A theory of justice. Revised*. Cambridge: Belknap Press.
Rubinstein, A. (1982). Perfect equilibrium in a bargaining model. *Econometrica, 50*, 97–110.
Schelling, T. (1959). For the abandonment of symmetry in game theory. *The Review of Economics and Statistics, 41*, 213–224.
Schelling, T. (1960). *The strategy of conflict*. Harvard: Harvard University Press.
Schotter, A., & Sopher, B. (2003). Social learning and coordination conventions in intergenerational games: An experimental study. *Journal of Political Economy, 111*, 498–529.
Shubik, M. (1959). *Strategy and market structure: Competition, oligopoly, and the theory of games*. New York: Wiley.
von Neumann, J., & Morgenstern, O. (2007). *Theory of Games and Economic Behavior* (Commemorative edition). Princeton: Princeton University Press.
Young, H. P. (1995). *Equity*. Princeton: Princeton University Press.