EDITED BY

Kenneth J. Arrow
Stanford University

Robert H. Mnookin
Harvard University

Lee Ross
Stanford University

Amos Tversky
Stanford University

Robert B. Wilson
Stanford University

Barriers to Conflict Resolution

W. W. NORTON & COMPANY
New York • London

CHAPTER SEVEN

On the Interpretation of Two Theoretical Models of Bargaining

Ariel Rubinstein

INTRODUCTION

This chapter is devoted to a discussion of the interpretation of two game-theoretic models of bargaining: Nash's bargaining model and the alternating offers model. The standard image of the ideal game-theoretic model of bargaining is one which gives a clear prediction of the outcome in a wide range of bargaining situations (i.e., situations in which there exists a common interest to reach an agreement and a conflict of interest as to its content). Personally, I have never understood how a game-theoretic model of bargaining, even when it attains a "clear-cut result," can be viewed as a functional prediction of a bargaining outcome in a specific scenario; neither have I ever understood how such a model can provide a bargainer with tips on how to bargain. Therefore, I believe that theoretical models of bargaining need better justification and interpretation.

The need to interpret bargaining models is no different from the need to interpret other models in game theory, where questions and doubts are raised. One expects the theory to provide useful tools for achieving concrete goals, testable predictions of the outcomes of game situations, and even tips on how to play games. But the predictions offered by game theory are not comparable to those offered by the natural sciences; nor do they provide advice on how to play games. The usefulness of game theory is therefore questioned.

Bargaining theory is a convenient test case for this inquiry since in contrast

---

Numerous discussions with friends inspired me to the ideas included in this paper. I would especially like to thank Asher Wolinsky, who forced me to provide response to his perpetual questioning of my approach, and Hugo Sonnenschein, whose criticism of the alternating offers model in his presidential address to the Econometric Society presented me with a challenge I cheerfully accepted. I am also grateful to Peter Barsoon, Zvika Safra and Bob Wilson for their very helpful comments.

to most other areas of game theory, it provides several models that produce clear predictions. This allows us to focus on the content of the prediction rather than on the interpretation of its indeterminacy.

The approach taken in this paper takes the criterion for a bargaining resolution to be its being protected against certain types of objections which have simple and intuitive verbal meanings. In this respect, the approach here is close to that of cooperative game theory. The paper aims at provoking a shift of interest from the functional forms of the solution concepts to the verbal forms of the objections.

First, let us review Nash's model, the most fundamental and important model in bargaining theory.

NASH'S BARGAINING SOLUTION: A REVIEW

Nash's bargaining solution is a theory which looks for a sharp prediction of the bargaining outcome based on the bargainers' preferences defined over the set of possible agreements and their attitudes toward risk. The theory is organized around two concepts: the *bargaining problem* and the *bargaining solution*. The description of a Nash bargaining problem includes the elements which are conceived by Nash's theory to be relevant: a set of feasible agreements, an event called "disagreement," the preferences held by the bargainers, and their attitudes toward risk.

According to Nash, "risk" appears in bargaining theory through the possibility of a breakdown in negotiations. Therefore, preferences are taken to be those over the set of lotteries whose "certain prizes" are the agreements and the disagreement event.

Formally, a *Nash bargaining problem* is defined as a four-tuple $<X,D,\succsim_1,\succsim_2>$, where X is a set of feasible alternatives, D is the disagreement event, and $\succsim_1$ and $\succsim_2$ are preferences defined in the space of lotteries in which the prizes are the elements in $X \cup D$.

The above definition of the Nash bargaining problem (see Rubinstein, Safra and Thomson 1992) is different from the conventional presentation of his bargaining theory (see Nash 1950), which defines a bargaining problem in its "reduced form," $<S,d>$ where S is a set of pairs of numbers and d is a specified element in S. Any element in S is thought of as a pair of *von Neumann-Morgenstern utilities* derived from some feasible agreement. Formally, the condensed model is derived from the more detailed model through the representation of the preferences by their VNM utility functions, u_1 and u_2, and by taking $S=\{(u_1(x),u_2(x)) \mid x \in X\}$, the set of all pairs of utilities which can be derived from the elements in X, taking $d=(u_1(D), u_2(D))$. Implicit in this identification is the assumption that nothing in the physical description of the set of agreements is relevant to the bargaining

outcome. A certain pair $<S,d>$ can be derived from *many* different quadruples $<X,D,\succsim_1,\succsim_2>$. In particular, Nash's space of problems $<S,d>$ is spanned either by varying preferences over a fixed set of alternatives or by keeping preferences on a universe of possible alternatives fixed, but varying the set of alternatives. Nash's hidden assumption is that all problems that produce the same pair $<S,d>$ have the same payoff vector outcome.

Nash's theory imposes some restrictions on what is a "bargaining problem." To avoid degeneracy, it is required that there is at least one agreement which Pareto-dominates the disagreement event. Since the agreements differ only through preferences, any two agreements for which the two players are indifferent are indistinguishable; it is thus convenient to assume that there are no such pairs of agreements. The set X is assumed to be compact and the preferences are continuous. Most importantly, it is assumed that a bargaining problem is convex in the sense that for all $x,y\epsilon X$ and for all $\alpha\epsilon[0,1]$ there is $z\epsilon X$ such that both players are indifferent between z and $\alpha\cdot x+(1-\alpha)\cdot y$, the lottery which gives x with probability α and y with probability $1-\alpha$. Denote by $\alpha\cdot x$ the lottery which gives x with probability α and gives the disagreement event D with probability $1-\alpha$, i.e., $\alpha\cdot x=\alpha\cdot x+(1-\alpha)\cdot D$.

The second central concept in Nash's theory is that of a "bargaining solution." A *bargaining solution* is a *function* which assigns a *unique* agreement to every bargaining problem. Thus, a bargaining solution tries to give a prediction of the bargaining outcome to each of the problems in its domain.

The standard definition of the Nash bargaining solution is the agreement which maximizes the product $[u_1(x)-u_1(D)][u_2(x)-u_2(D)]$ over all agreements which Pareto-dominate the disagreement event. For the bargaining problems under consideration, the above formula defines a unique prediction. Nash (1950) not only proposed this solution concept, but also showed that this solution is unique in satisfying the following three postulates: *symmetry* (SYM), *Pareto optimality* (PAR) and *independence of irrelevant alternatives* (IIA). Nash's fourth axiom, the *invariance to positive affine transformations* (INV), is redundant in the current setting, since we start from the players' attitudes toward risk and not from utility representations.

The first axiom is easy to accept. It captures the idea that the prediction rests only on the information included in the bargaining situation. A bargaining problem is called symmetric if players 1 and 2 are "indistinguishable" (presented in its reduced form $<S,d>$, a bargaining problem is symmetric if the set S is symmetric with respect to the main diagonal and d is a point on the main diagonal; for a definition in terms of preferences, see Rubinstein, Safra, and Thomson 1992). SYM requires that for symmetric problems, the predicted agreement does not discriminate between the two players (in utility terms, the solution concept should assign equal utilities to both players).

PAR requires that, for any bargaining problem, there is no agreement in X which Pareto-dominates the solution outcome. The axiom excludes the possibility that bargaining will result in disagreement.

The most problematic axiom is the IIA: its standard definition is that "if $X \subseteq X'$ and the solution to the problem $<X',D,\succsim_1,\succsim_2>$ is in X, then it is also the solution to the problem $<X',D,\succsim_1,\succsim_2>$". An alternative formulation of the IIA is suggested in Rubinstein, Safra, and Thomson 1992. Assume that y^* is the solution to the problem $<X,D,\succsim_1,\succsim_2>$ and let $\succsim_i'$ be a preference which agrees with $\succsim_i$ on the set of deterministic agreements X such that:

1. for all x such that $x \succsim_i y^*$, if $p \cdot x \sim_i y^*$ then $p \cdot x \precsim_i' y^*$, and
2. for all x such that $x \precsim_i y^*$, if $x \sim_i q \cdot y^*$ then $x \sim_i' q \cdot y^*$.

Then, y^* is also the solution to the problem $<X',D,\succsim_i',\succsim_j>$. The switch of agent i's preference from $\succsim_i$ to $\succsim_i'$ reflects his increased aversion to the risk of demanding alternatives which are better than the outcome y^*. Though player i still prefers x to y^*, he is less willing to risk demanding x. The axiom captures an intuition that the bargaining solution outcome y^* should be defensible against possible objections. The change in player i's preference described in the axiom makes player i "less eager" to object. The axiom states that this change does not affect the bargaining outcome.

THE INTERPRETATION OF THE NASH BARGAINING SOLUTION

One of the main reasons for the popularity of the Nash bargaining solution among economists is that it is defined by a very simple formula which is easily embedded in any larger model that includes a bargaining component. The analytical convenience is an important reason to use it in economic models. If the task of bargaining theory is to provide a "clear-cut" prediction for a wide range of bargaining problems, then Nash bargaining theory carries out the task perfectly well. The Nash bargaining solution is well defined and the axioms lend the solution a sense of non-arbitrariness for a wide range of problems. However, I cannot see the formula being tested as similar simple formulae in sciences are. In the absence of testability, we search for a meaning for the formula. What is the product of two von Neumann-Morgenstern utility numbers and what is the meaning of the maximization of that product? Can we consider the maximization of a product of utilities as an intelligent principle for resolving conflicts? The negative answer to these questions prompts us to look for alternative definitions of the Nash bargaining solution. Such a definition is required to have an attractive verbal (as opposed to analytical) meaning in ordinary language. It has to be defined by a simple sen-

tence which includes only the terms "alternative," "disagreement," and "preference."

One may argue that the conjunction of the Nash axioms is an attractive alternative definition for the Nash bargaining solution. The axioms are defined in quite nontechnical terms and the Nash theorem can be interpreted as a proof that the conjunction of the axioms provides an implicit definition of the Nash bargaining solution. However, this is an implicit definition and we are looking for an explicit definition, one which will specify the outcome of a particular problem directly in terms of the problem's primitives without the need to search for consistency with the outcomes assigned by the solution for other problems.

I would now like to propose such an alternative definition (see Rubinstein, Safra, and Thomson 1992; the idea is close in spirit to Zeuthen's [1930] though it is quite different).

Definition: The agreement y^* is the ordinal Nash solution for the problem if and only if for any player i, any agreement $x \in X$ and any probability $p \in [0,1]$, if $p \cdot y^* \prec_j x$ then $p \cdot x \prec_i y^*$.

Thus, a Nash bargaining solution is an agreement y^* such that for any objection by player i, who proposes x rather than y^* and who takes steps which may cause disagreement with probability $1-p$, either

(1) it is credible for player j to reject the objection and to insist on the agreement y^*, even when he takes into consideration the possibility of breakdown, or,
(2) it is not credible for j to reject the objection ($p \cdot y^* \prec_j x$) but player i prefers not to take the risk ($y^* \succ_i p \cdot x$).

In other words, the Nash bargaining solution is the agreement such that any argument of the type "You should agree to my request, x, since x is better for you than insisting on the convention y^* given the probability $1-p$ of breakdown" is not profitable to the objector when he takes into account the same probability of breakdown.

The Gulf war provides a concrete example of this definition. The bargainers were Iraq and the U.S. The set of agreements contained the various possible partitions of the land in that region. The disagreement event was a war. When Saddam Hussein moved his troops he deliberately took a chance that the situation would deteriorate into an unpleasant war before the U.S. gave up. He preferred the lottery in which he would find himself in a war with a probability $1-p$ and that he would annex Kuwait with the probability p to the alternative of maintaining the status quo. However, his mistake was that the U.S. preferred the risk of war in demanding a return to the status quo rather than giving in to Iraq's demands. If the U.S. had given in to Iraq it would have meant that the pre-invasion borders were not part of a Nash bargaining outcome.

The first thing we notice about the ordinal definition is that in the expected utility case it leads to the same outcome as does the conventional definition of the Nash solution. Without loss of generality, we can choose the utility representation so that for both players $u_i(D)=0$. Then, the alternative y^* satisfies $u_1(y^*)u_2(y^*) \geq u_1(x)u_2(x)$ for all x for both i *if and only if* for all i, for all $x \succ_i y^*$ for all $p \leq 1$, if $u_j(x)/u_j(y^*) > p$ then $u_i(y^*)/u_i(x) > p$ *if and only if* for all i and x, $py^* \prec_j x$ implies $px \prec_i y^*$.

In recent years there has been a growing interest in nonexpected utility theories of decision making under uncertainty since they explain a wide range of behavior patterns and experimental results which are inconsistent with expected utility theory. The ordinal definition above uses the language of agreements and preferences without a reference to utilities, thus it has the advantage that it can be extended to bargaining situations in which the preferences do not satisfy expected utility theory conditions.

The above explanation of the Nash bargaining solution also has a normative version. One can say that the Nash bargaining solution is an agreement in which no player i can make the following claim against player j: "You should agree to x since I prefer x to y^* so much that I am ready to take the $(1-p)$ probability risk of breakdown while you are not."

Notice also that in the (ordinal) definition of the Nash bargaining solution, there is a symmetry in the two bargainers' beliefs concerning the possibility of a breakdown. One can imagine other scenarios in which there is a systematic asymmetry in the beliefs about the possibility of breakdown. If we refine the definition of the Nash bargaining solution so that the probability p appears in player 1's considerations and is not the same as the probability p that appears in player 2's considerations, then we can arrive at a solution called an "asymmetric Nash bargaining solution."

THE ALTERNATING OFFERS MODEL: A REVIEW

This section is a brief review of the infinite-horizon alternating offers model (see Rubinstein 1982). Actually, I will present a variation of the model (see Binmore, Rubinstein, and Wolinsky 1986) in which the primitives of the model are the four-tuple $<X, D, \succsim_1, \succsim_2>$ as in the Nash model. The model is built upon a specific procedure used to reach agreement in which the players alternate turns in having the right to make offers. In each period, one of the players must make a proposal and the other must either accept or reject it. Acceptance ends the game. In the case of rejection, the responder has the right to make a proposal. However, before he does so, "nature" may interfere and cease the game, causing the outcome D. The probability of breakdown is fixed, $1-p>0$. One interpretation of fixing $(1-p)$ is that the risk of breakdown is a function of the length of the time

interval between a rejection and a counteroffer; the equal probability of breakdown reflects the assumption of the equal length of time between responses and counteroffers.

The above procedure, together with the preferences, forms an extensive game. Under the assumptions made on the bargaining problem in section 2, there is a unique subgame perfect equilibrium which is characterized by the unique pair of Pareto-optimal agreements x^* and y^*, so that player 1 is indifferent between y^* and the lottery $p \cdot x^*$ and player 2 is indifferent between x^* and the lottery $p \cdot y^*$. The following are subgame perfect equilibrium strategies: player 1 always makes the offer x^* and accepts any offer as good as y^*. Player 2 always makes the offer y^* and accepts any offer as good as x^*.

The current version of the alternating offers model has an interesting connection with the Nash bargaining solution which was discovered by Binmore (1987) and clarified by Binmore, Rubinstein, and Wolinsky (1986). Let $x^*(p)$ and $y^*(p)$ satisfy $p \cdot x^*(p) \sim_1 y^*(p)$ and $p \cdot y^*(p) \sim_2 x^*(p)$. Where $p \to 1$, both $x^*(p)$ and $y^*(p)$ converge to the Nash bargaining solution of the problem $<X, D, \succsim_1, \succsim_2>$. In other words, where the probability of breakdown is very small, the agreement reached is very close to the Nash bargaining solution, independent of which player makes the first proposal.

As in the case of the Nash bargaining solution, the above analysis achieves the standard objectives of game theory. Here we have quite a natural model which has a unique subgame perfect equilibrium and whose outcome is a clear-cut prediction of the outcome of bargaining.

By using the noncooperative approach, we have added a procedure to the description of the bargaining problem. The set of subgame perfect-equilibrium outcomes is sensitive to the procedure of bargaining. The sensitivity of the model to the procedure is an advantage of the noncooperative approach and one reason for its popularity in economic theory, as it allows for the modeling of trading institutions.

An agreement in a bargaining problem is taken in this paper to be stable if no bargainer can raise a *valid* argument which is *worthy* of raising. The contents of the term "valid" and "worthy of raising" differentiate the two models discussed in this paper. In the interpretation of the Nash bargaining solution, raising an objection z to y* is accompanied with creating conditions so that disagreement may happen. A valid argument is a statement of the type: "you better agree to z; for you, insisting on y* does not merit taking the risk of breakdown". The reference point in this type of argument is the current agreement y*. As to the objector, he assesses the value of raising an objection by taking into account that the risk of breakdown applies to him as well.

In the alternating offers model, on the other hand, the risk of breakdown is endogenous to the situation. As long as he believes that the partner will

accept his objection, an objector does not take into account any risk when he raises an objection. The structure of a valid argument is "you better agree to z since anything I would accept is worse for you given the risk you face if you reject the offer". Thus, the reference point of the responder is the acceptance set of his opponent.

Other modifications of the alternating offers procedure lead to less attractive conclusions. If player 2 is allowed to make offers only once in three periods, the only subgame perfect equilibrium outcome is close to the partition (2/3,1/3) (when $1-p$ is small). When the probability of breakdown is small, it seems strange that such a relatively small change in the procedure should make such a significant impact on the bargaining outcome. This observation, together with the fact that real-life bargaining only rarely has a rigid procedure, was the main source of criticism against the alternating offers model (see for example Kreps 1990b and Sonnenschein 1991).

AN ALTERNATIVE INTERPRETATION OF THE ALTERNATING OFFERS MODEL

The above criticism is derived from the classical interpretation of the game form as a full description of the physical events in the modeled situation. In contrast I wish to follow Nash's approach: "Of course one cannot represent all possible bargaining devices as moves in the non-cooperative game. The negotiation process must be formalized and restricted, but in such a way that each participant is still able to utilize all the essential strengths of his position" (Nash 1953, 129). I view a game as a description of the *relevant* factors involved in a specific situation as perceived by the players rather than as a presentation of the physical rules of the game. According to this understanding, the alternating offers model is a model which captures the interaction between two bargainers whose reasoning is using choices between an offer on the table and a possibility to achieve a better agreement if they take a certain risk of a breakdown of negotiations. Thus, if the alternating offers model is interesting, it is not because it describes a real-life bargaining process but because it embeds an interesting type of consideration which players use in bargaining. Thus, a variant of the model in which the probabilities of breakdown after rejection by player 1 and 2 are distinct captures an asymmetry in player evaluations of breakdown. On the other hand, the version of the model, suggested by the critics, in which player 1 makes offers only once during three periods, fails to capture any sensible consideration. In the rest of this section, I will suggest a somewhat different model, one that allows us to analyze the strategic considerations which appear in the alternating offers model without using noncooperative games.

The new model is based on a view of a social order as an automaton. The automaton consists of the following elements: a set, S, interpreted as the set of "states of the system" and an initial state, s_0. Each state is accompanied by two sets of agreements $A_1(s)$ and $A_2(s)$, with the interpretation that at state s offers in the set $A_i(s)$ made by player i are expected to gain acceptance. (In the automatae jargon, this is the output function of the automaton). The automaton responds to events in the bargaining session. In the present context, the events are offers and responses; in alternative models we could also include other types of actions as events (walking away from the table, making commitments, etc.). The transition function describes the moves from one state to another as a function of the possible events in the game.

For a social order to be stable, we require that:

1. For each s and for every $b \notin A_i(s)$, the system moves to state s' after player i makes an offer b and there is an agreement $x \epsilon A_j(s')$, so that $p \cdot x \gtrsim_j b$. (Note that p is here a constant). In other words, if b is not acceptable, there must be an acceptable counteroffer made by player j, so that it is optimal for player j to reject b and to insist on x even when he takes the risk of breakdown into account.
2. If $a \epsilon A_i(s)$, then if i offers a and j rejects a, the system moves to state s' so that there is no $b \epsilon A_j(s')$ that satisfies $pb \succ_i a$.

It can be shown that the only automaton that satisfies the above conditions has the property that for any s, $A_1(s) = \{x | x \lesssim_1 x^*\}$ and $A_2(s) = \{x | x \lesssim_2 y^*\}$.

In light of this result, which characterizes the acceptance sets, a convention must be included in the set $\{x | x \gtrsim_1 y^* \text{ and } x \gtrsim_2 x^*\}$ in order to be stable in face of acceptable objections. Consider, for example, an agreement x, so that $x \prec_1 y^*$. It is profitable for player 1 to raise the acceptable objection x^*, because he prefers the lottery px^* over the agreement x. Notice that here, the response to an unjustified demand is not to insist on the status quo (such as in the interpretation of Nash's bargaining solution) but to raise a counterdemand which has to be acceptable. A bargainer may be deterred from starting negotiations (rather than maintaining the status quo) by his anticipation of a counterdemand which is worse for him than the status quo.

The equivalence between the above model and the alternating offers model results from the fact that the notion of subgame perfect equilibrium in the alternating offers model has the following "one deviation property": a pair of strategies is a subgame perfect equilibrium if and only if there is no history after which a player can gain by a one-time deviation.

The advantage of the new concept lies in the fact that it models strategic considerations directly and avoids the need to use the problematic notion of an "extensive game strategy," which requires a player to make plans for action after an unbounded number of times in which he has not followed his strategy (see Rubinstein 1991).

FINAL DISCUSSION

My aim in this paper has been to shift the focus of theoretical bargaining models from "formulae" to "argumentation." Casual observation of real-life negotiations indicates that seldom does a bargainer simply make an offer. An offer is usually accompanied with arguments. The arguments may be concerned with the underlying interests ("If you disagree, you may lose the opportunity to make a deal") or with fairness concepts ("I gave up more than you did") or may just be rhetorical ("A taxi driver offered me a discount the size of the tolls on the fare"). A bargaining solution is taken to be an agreement for which no acceptable objection exists, given the arsenal of arguments available to the bargainers.

In this chapter, I have dealt with only one type of argument: "You should agree to x; if you demand the admittedly acceptable y you take the risk of breakdown and this is not worthwhile for you." The Nash bargaining solution is the agreement that survives this argument if the players identify what is acceptable after raising objections with the convention. All agreements between $x^*(p)$ and $y^*(p)$ are those for which any objection is credibly rejected by a counteroffer which is acceptable in an internally consistent sense and given that p is a fixed probability of breakdown whenever an offer is rejected.

This paper has dealt with the interpretation of old concepts and results. I believe that the approach presented in this paper can provide grounds also for new results. But like any other statement of this sort, its proof is only in the doing.