The Center for Mathematical Studies in Economics and Management Science
Northwestern University, Evanston, Illinois 60201

Discussion Paper No. 658

NEGOTIATION IN GAMES: A THEORETICAL OVERVIEW

by

Roger B. Myerson

July 1985

Acknowledgements. This paper was written with the support of a grant from the Alfred P. Sloan Foundation. Helpful suggestions by Peter Cramton are also gratefully acknowledged.

This paper was prepared for presentation at the I.M.S.S.S. summer workshop at Stanford University.

NEGOTIATION IN GAMES: A THEORETICAL OVERVIEW

by Roger B. Myerson

1. The perspective of cooperative game theory

The goal of this paper is to offer a general perspective on what the methods of cooperative game theory can tell us about bargaining and negotiation between individuals who have different private information. The logical foundations and methodology of cooperative game theory may not be as clear-cut as those of noncooperative game theory, but cooperative game theory does provide a conceptual structure that can give important insights into the practical problems of negotiation. In particular, the analysis of cooperative games with incomplete information can explain how regrettable (or ex-post inefficient) outcomes, such as strikes, costly delays, and litigation, can occur in an efficiently designed social system, which older theories of cooperative games with complete information could not explain. This paper will try to survey the basic ideas and results of the theory of cooperative games with incomplete information at a conceptual level, without getting deeply into the technical detail. For a more detailed and broader introduction to game theory, see Myerson [1984d] and [1985a].

The fundamental principle of game theory is that any definitive theory or social plan that predicts or prescribes behavior for all players in a game must designate a Nash equilibrium (or, more precisely, a sequentially rational Nash equilibrium), if this theory or plan is to be understood by all the players and is not to impute irrational behavior to any player at any point in time. (Here, by "sequentally rational", we mean that the equilibrium should satisfy some conditions like those discussed by Kreps and Wilson [1982] or

Myerson [1984c], or at least the subgame-perfectness condition of Selten [1975].) Thus, the basic method of "noncooperative" game-theoretic analysis is to formulate an extensive or strategic form model for the social situation in question, and then identify all (sequentially rational) equilibria of this model.

There are two general problems with such analysis: there may only be dismal equilibria (as in the well-known "prisoner's dilemma" example), or there may be multiple equilibria (as in the well-known "battle of sexes" example, see Luce and Raiffa [1957]).

The general response to dismal-equilibrium problem is, if possible, to make a cooperative transformation of the game, by adding communication or contract-signing options, or by repeating the game, so as to create a new game which may have equilibria that have better welfare properties. Mediators, regulators, and auditors are outside interveners who assist in such cooperative transformations. A mediator is any person or machine that can help the players to communicate. The revelation principle asserts that a central mediator subject to incentive constraints can simulate any equilibrium of any communication-transformed game. The effect of a regulator or auditor is to relax some of these incentive constraints. A regulator helps the players to precommit to do some future action that they might not actually want to do when the time of action comes, and an auditor helps the players to avoid the temptation to lie about their private information.

Nash's [1951] program for the analysis of bargaining is to formulate and analyze the cooperatively-transformed game by noncooperative methods. However, these cooperative transformations usually exacerbate the multiple-equilibrium problem. For example, in the demand game formulated by

Nash [1953], every physically feasible allocation that gives each player more than his disagreement payoff is the outcome in some equilibrium. For the purposes of a social planner or arbitrator who can select among the available equilibria, such multiplicity may be good news, but for the purposes of a theorist who wants to make tight predictions, such multiplicity is very bad.

The basic response to the multiple-equilibrium problem is Schelling's [1960] focal point effect. This asserts that, in a game with multiple equilibria, anything that tends to focus the players' attention on one particular equilibrium, in a way that is commonly recognized, will tend (like a self-fulfilling prophecy) to make this the equilibrium that the players actually implement. Criteria for determining a focal equilibrium may be grouped into seven general catagories as follows.

(1) Environmental factors and historical precedent may determine the focal equilibrium that the players implement. To the extent that such factors are important and depend on structures that are ignored in the basic models of game theory, their effect is to limit the power of game-theoretical analysis.

(2) Strategic properties such as simplicity or stationarity may make some equilibria seem more natural and hence focal.

(3) Being the limit of equilibria of some sequence of perturbed games that converges to the actual game may make some equilibrium focal. A basic difficulty with this kind of criterion is that the limits of equilibria may depend on which sequence of perturbed games is considered. Nevertheless, this idea was used in various forms by Nash [1953] and Binmore [1981] to provide one justification of Nash's bargaining solution. This idea is also part of the motivation for concepts of stable equilibria defined by Kohlberg and Mertens [1983] and for Harsanyi's [1975] tracing procedure.

(4) Judgement of an outside *arbitrator* or social planner may determine the focal equilibrium. That is, we may say that an arbitrator is any outside individual who has the authority or power of persuasion to make an equilibrium focal by advocating it in some pre-play communication.

(5) Objective standards of *equity* and *efficiency*, which might be the basis of an impartial arbitrator's judgement, may determine the focal equilibrium even when no arbitrator is present, if the players all understand these standards and can therefore predict the equilibrium that such an impartial arbitrator would have selected.

(6) There may be one player in the game, whom we may call the *principal*, who has the arbitrative power to to make an equilibrium focal by advocating it in pre-play communication.

(7) The focal equilibrium may be determined by a consensus reached in *negotiations* that involve some or all of the players. That is, for a given game, we may define a *negotiation* to be any processs of pre-play communication among the players, which involves no actions that are intrinsically relevant to payoffs, but which can influence the outcome of the game by determining a focal equilibrium.

To relate (4), (6), and (7) above, we could say that a principal, in our terminology, is a player who can act like an arbitrator, or, equivalently, is a player who has all of the negotiating ability.

One could try to analyze processes of arbitration and negotiation by making them an explicit part of an (augmented) extensive-form model of the game. To do so, however, is to represent a process for selecting equilibria in one game as a part of an equilibrium in a larger game, which generally may have an even larger set of equilibria, so that we are still left with an equilibrium-selection problem. For example, pre-play announcements by an arbitrator or principal could be included as the first stage in the augmented game model (where the set of permissible announcements may be identified with the set of sequentially rational equilibria of the original game), but there would always be sequential equilibria of this augmented game in which the players all ignore this initial announcement. Of course, there will also be sequential equilibria in this augmented game in which the players always implement the announced equilibrium, but the traditional methods of noncooperative game theory have not offered any criterion to lead theorists to focus on these.

Thus, our analysis of (4), (6), or (7) must be directly based on some further assumption that expresses the idea that the arbitration or negotiation process will effectively influence the equilibrium perceptions of all players in the direction that the arbitrator, principal, or negotiating players desire. The significance of such an effectiveness assumption must be that the selected equilibrium will systematically depend on the preferences of the relevant individuals (arbitrator, principal, or all the negotiating players) over the set of all equilibria, in some way such that the more preferred equilibria are more likely to occur. If players can negotiate effectively then they should be expected to play an equilibrium which is not Pareto-dominated for them within the set of possible equilibria. Furthermore, an effectively negotiated equilibrium should give to each player an expected payoff that is in some sense commensurate with the relative strength of his position in the game (what he has to offer others, what he can guarantee himself) and his negotiating ability.

Criteria (5), (6), and (7) all depend on some kind of comparison of the welfare properties (that is, the players' expected payoff allocations) of the various equilibria. The basic goal of cooperative game theory is to develop a formal theory to predict the outcome of cooperatively transformed games in which any of these welfare-based focal criteria are effective.

2. Complications in the extending the theory of equilibrium selection

The idea that the focal equilibium may be selected by one individual, or negotiated by several individuals, is important and relevant to many real situations; but it also creates some subtle theoretical difficulties, which any general theory of equilibrium selection will need to resolve.

The significance of assuming that some principal player can select the focal equilibrium seems clear, provided that he makes this selection before learning any private information. In this case, the (sequentially rational) equilibrium that gives the highest expected utility to the principal should be implemented. However, if the principal has private information at the time of equilibrium selection, then he may face a dilemma. His preferred equilibrium may depend on his information, but may cease to be an equilibrium if its selection depends on (and thus reveals) this information. To resolve this dilemma, the principal's selection must reflect an inscrutable compromise between his possible preferences. We will discuss this issue further in Section 4.

There is also some conceptual difficulty if a principal has arbitrative power only over a subset of the players or stages of the game. To see this, for example, let us consider the following two two-stage games. In the first game, at stage 1, there is a chance move that chooses "H" or "T" with equal probability, and then, at stage 2, player 1 chooses "h" or "t" without observing the move from stage 1. Player 1 gets a payoff of +1 if he matches the chance move and -1 if he does not. Player 2 has no intrinsically payoff-relevant moves (nor does he observe the chance move) but he has the arbitrative power to select among the multiple equilibria of this game, and he prefers that 2 chooses "h". Then we should expect that 1 will be persuaded by 2 to choose "h". In the second game, the initial move is instead chosen by some player 3, whose payoffs are opposite to 1's, rather than by chance. Suppose furthermore that 2 still has the same persuasive power to influence 1 in this second game, but that 3's choice is made before 2 can say anything. Then, although the relationship between players 1 and 2 in stage 2 is exactly the same in both games, and player 3 must in equilibrium be randomizing .50-.50 like the chance move in the first game, we must conclude that player 2 will not be able to persuade player 1 to choose "h" for sure in the second game, because 2 randomizes in the unique equilibrium. This suggests that some care is needed to define solutions to sequential principal-agent problems, where the principal will retain his persuasive power to influence equilibrium expectations in later stages of the game. Something like the *principal's quasi-equilibrium* of Myerson [1982] may be needed to resolve this difficulty. We will return to this issue in Section 7.

The assumption that players can negotiate effectively also should apply to subcoalitions as well as to the grand coalition of all players. To see this, consider the two-player "divide the dollars" game with a dummy third player. We normally expect that the two nondummy players will coordinate on an allocation that divides all the money between themselves, because their subcoalition could negotiate effectively to block any allocation in which they give money to the dummy. However, the "three-player majority game" (where any allocation of the available money can be implemented if it is approved by a majority of the three players), or any other game with an empty core, shows that the ability of one subcoalition to negotiate effectively must be in some sense limited by the ability of other coalitions to negotiate effectively. The Shapley value and other characteristic-function solution concepts are attempts to describe how the conflicting abilities of various subcoalitions of players to negotiate effectively among themselves should be reconciled.

3. The equity hypothesis.

To begin an analysis of negotiated outcomes of games, we may use the following equity hypothesis: the reasonable outcomes of effective negotiations in which the players have equal opportunity to participate should be the same as the reasonable recommendations that would be made by an impartial arbitrator who has the same information as is common knowledge among the players during the negotiations. On the one hand, an impartial arbitrator should not steer the players to something that could be worse for some players than the outcome that they would have negotiated without him, if this outcome is known. On the other hand, if it is obvious what an imp_rtial arbitrator would have suggested, then the most persuasive argument in negotiations should be to settle on this equilibrium (see, for example, Fisher and Ury [1981]). Since we have not yet rigorously defined "symmetric effective negotiations" or "impartial arbitration," this equity hypothesis cannot be derived as a formal game-theoretic theorem. It is instead an assertion about two different intuitive concepts, both of which we want to formalize. The power of the equity hypothesis comes from the fact that it allows us to carry any restrictions that we can make on one concept over to the other.

For example, in the "divide the dollars" game (where two risk-neutral players can divide $100 if they can agree on its division, otherwise both get zero), we may predict the $50-$50 division as the outcome of a negotiated settlement, because it is obviously the solution of an impartial arbitrator. More generally, the equity hpothesis suggests that the outcome of negotiations between players of symmetric negotiating ability should be an agreement that is in some sense equitable and efficient. That is, the equity hypothesis suggests an equivalence between the focal criteria (5) and (7) in Section 1.

Nash's [1950] axioms for two-person cooperatiove games may be viewed as normative restrictions on the way that an ideal impartial arbitrator should behave. (The arbitrator's recommendations should be Pareto-efficient and individually rational, should treat players symmetrically if their positions are symmetric in the feasible set and the disagreement outcome that would occur if all arbitration and negotiation failed, should depend only on the underlying decision-theoretic properties of the players' preferences, should not be based on any distinction among outcomes between which both players are indifferent, and should be independent of the elimination of options that would not be chosen with or without his arbitration.) Then the equity hypothesis translates these normative assumptions into a positive prediction that the outcome of negotiations between two players should maximize the poduct of their expected utility gains over the disagreement outcome.

On the other hand, we may have little intuition about how an impartial arbitrator should compromise between the conflicting interests of two different possible types of one player, in a game of incomplete information. Thus, an analysis of the requirements for inscrutability in the case of negotiated settlements may help to determine what are reasonable arbitrated agreements for such a game.

4. Fundamental isssues of cooperation under uncertainty.

We now consider a series of basic propositions which should underlie any theory of cooperation under uncertainty.

(1) As developed by Harsanyi [1967-8] (see also Mertens and Zamir [1985] and Myerson [1985a]), the Bayesian game model is the appropriate basic model for describing games in which the players already have different information at the beginning of play, that is, games with incomplete inforamtion. To specify a Bayesian game, we first specify a set of players N. Then, for each player i in N, we must specify the set of i's possible actions C_i, and the set of i's possible types T_i. A type for i is any possible specification of all of i's private information relevant to the game. For each player i, we specify a utility function u_i that determines how i's payoff (measured in some von Neumann-Morgenstern utility scale) depends on the actions and types of all players. Finally,for each i, we specify a probability function p_i that determines the probabilities that i would assign to each possible combination of types for the other players, as a function of his own type. These structures of the Bayesian game model $(N,(C_i,T_i,u_i,p_i)_{i\in N})$ are supposed to describe all of the information that is common knowledge among the players at the beginning of the game. The private information of each player i is represented by his type, which is some element of the set T_i.

(2) In studying a Bayesian game, the appropriate object for welfare analysis by any outside arbitrator must be the mechanism (or decision rule) that determines the actions taken, rather than the actions themselves. Here, a mechanism is any rule for randomly determining the players' actions as a function of their types. That is, a mechanism is a function μ that specifies a probability distriution over the set of possible combinations of actions $(\times_{i\in N} C_i)$ for every possible combination of types in $\times_{i\in N} T_i$. An outsider can hope to influence the process by which the players choose their actions as a function of their types, but he cannot influence their given types. Since the actual types are also unknown to any outsider, he can only judge the impact of his potential influence on the players by how he changes the mechanism by which they will choose their final actions, because he generally cannot predict what those final actions actually will be without knowing their types.

(3) A mechanism is *incentive compatible* if it could be implemented by a central mediator who communicates directly and confidentially with each player in such a way that no player would have any incentive to lie to the mediator or disobey his recommendations, when all other players are expected to be honest and obedient. When the basic Bayesian game is transformed by adding any kind of opportunities for communication among the players, any equilibrium of this transformed game must be equivalent to some incentive-compatible mechanism. (This is the *revelation principle*.) These incentive-compatible mechanisms can be formally characterized by a list of mathematical incentive constraints, which assert that no player should expect to gain by dishonestly or disobediently manipulating the mechanism.

Thus, we should think of the set of incentive-compatible mechanisms as the set of feasible equilibria among which an arbitrator could choose. To see that choosing an incentive-compatible mechanism is the same as choosing among multiple equilibria, suppose that there are an infinite number of mediators, one for every incentive-compatible mechanism. Every mediator plans to receive a type-report from every player and then return an action-recommendation to every player, where the mediator's recommendation will be determined according to the mechanism that he represents. There are multiple equilibria of this game, which an arbitrator or negotiating players must choose among. Specifically, for any mediator, there is an equilibrium in which the players all send honest reports to him and obey his recommendations, and all players send meaningless type-independent reports to all other mediators, whose uninformative recommendations are then universally ignored.

(4) An incentive-compatible mechanism is _interim incentive-efficient_ (or simply _incentive-efficient_, for short), in the sense of Holmström and Myerson [1983] if there does not exist any other incentive-compatible mechanism that gives higher expected utility to some types of some players and lower expected utility to none. Let $U_i(\mu|t_i)$ denote the expected utility for player i if his type is t_i and the players actions are to be determined according to the mechanism μ. Then, μ is interim incentive-efficient iff μ is an incentive-compatible mechanism and there does not exist any other incentive-compatible mechanism ν such that $U_i(\nu|t_i) \geq U_i(\mu|t_i)$ for every type t_i of every player i, with at least one strict inequality. Interim incentive-efficiency is the basic welfare efficiency criterion that should be applied by any benevolent arbitrator. A mechanism is interim incentive-efficient if and only if it is not common knowledge that some other feasible mechanism would be weakly preferred by every player, given his private

information, and might be strictly preferred by some. That is, if a mechanism is not efficient in this sense, then there is a feasible change of mechanism that every player would approve, given his current information, and which might be strictly preferrable for some players.

(5) An incentive-efficient mechanism may generate outcomes that will not be Pareto-efficient ex post, after all players' types are revealed, because the incentive constraints that bind in the mechanism selection problem will not be evident ex post. For example, an incentive-efficient mechanism may need to allow for a positive probability of a strike when the firm claims that its ability to increase wages is surprisingly low, in order to satisfy the incentive constraint that the firm should not expect to gain by underrepresenting its ability to pay. So incentive-constraints can explain how regrettable outcomes may occur in a well-designed social system.

(6) Much economic analysis has been done under the assumption that the mechanism used in a cooperative game with incomplete information must be ex-ante incentive-efficient, in the sense that there is no other incentive-compatible mechanism that all players would have preferred (weakly, but with at least one strictly) before learning their types. This is a stronger criterion than interim incentive-efficiency, but the argument for it does not seem compelling when we assume that each player already knows his own t̬pe at the time that any negotiation or arbitration begins, so that calculations of ex ante (pre-type) expected utility are irrelevant to him. (We are assuming that this game is not to be repeated, or, more generally, that any relevant repetitions of the existing situation are already been included — in our definition of the game.) Thus, focusing on ex-ante efficiency may lead to some systematic biases in the analysis of cooperative games. For

example, it is easy to construct labor-management bargaining games in which costly underemployment (relative to the full-information optimum) occurs with positive probability in all interim incentive-efficient mechanisms except those which are ex-ante incentive-efficient. (There are other examples, however, in which even ex-ante incentive-efficent mechanisms have underemployment.)

(7) The equity hypothesis does not imply that arbitration and negotiation should be completely equivalent for cooperative games with incomplete information unless the players negotiate <u>inscrutably</u> (that is, without revealing anything about their private information during the negotiation process). Otherwise, the information that would become common knowledge during negotiations might be greater than the information that is common knowledge initially, which is all that an arbitrator could use in planning a mechanism. This issue is related to the distinction between efficiency and <u>durability</u> made by Holmström and Myerson [1983]. However, Holmström and Myerson were considering a social-choice environment in which the noncooperative actions available to the individual players were not specified, as Crawford [1985] has observed.

In general, to be able to negotiate inscrutably, a player must be able to phrase his negotiating strategy in terms of type-contingent mechanisms, rather than just in terms of final combinations of actions. That is, if players can agree on mechanisms, then they can agree to share information without actually doing so during the negotiation process. Using this idea, Myerson [1983] has argued that there is no loss of generality in assuming that a principal with private information will choose an incentive-compatible mechanism inscrutably. The same inscrutability assumption is also used, but less convincingly, in the analysis of bargaining problems where two or more players can negotiate, in Myerson [1984a,1984b].

In any case, any definitive theory of negotiated settlements must ultimately specify a reduced-form solution that is a single incentive-compatible mechanism for the game in question. If the players do not negotiate inscrutably, then this reduced-form solution is the composition of a settlement function which maps type-combinations into the agreements that may be reached at the end of the negotiations, and an interpretive function which determines the ultimate combination of actions in $\times_{i\in N} C_i$ as a function of the players' types and the agreeement reached.

(8) In general, there is an issue of intertype compromise as well as interpersonal compromise that must be determined in any theory of cooperative games with incomplete information. For example, there may be two different interim incentive-efficient mechanisms that give the same expected payoffs to all types of player 2, but one may be better for type A of player 1 while the other is better for type B of player 1. If one of these is the (arbitrated, negotiated, or principal-selected) solution for the cooperative game, then some tradeoff between payoffs to the two different potential types of player 1 must have somehow been made.

The most important basis for making interpersonal compromise is generally accepted to be the need to satisfy some kind of equity criterion (lest the subequitably-treated players reject the agreement). The natural basis for determining intertype compromise (between two types of the same player) in negotiation is the player's need to satisfy some inscrutability condition, lest the other players infer something about his type from his negotiating strategy and use this information against him. If type A of player 1 is not particularly concerned to conceal his type from player 2 (because A is the "good" type), then the inscrutable compromise will probably favor the preferences of type A. (If neither type of player 1 needs to conceal his information from player 2, then the situation described in the preceding paragraph probably would not arise.)

(9) To understand the issue of inscrutable intertype compromise, it is best to begin with the case of mechanism selection by a principal player, because this case is conceptually the simplest in the case of games with complete information. There is one important class of games in which we can define a clear resolution of the problem of compromise between the conflicting goals of the principal's types: these are the games in which the principal has a strong solution, in the sense of Myerson [1983]. A strong solution for the principal is an incentive-compatible mechanism that has two properties: it would still be incentive-compatible even if the principal's type were publicly revealed to all the other players, no matter what that type may be (so it is safe for the principal); and there is no other incentive-compatible mechanism that is weakly better for all of the principal's types and strictly better for some (so it is undominated for the principal). It can be shown that the strong solution must be essentially unique, if it exists. Because it is safe, the principal does not need to conceal his type when he advocates a strong solution. On the other hand, it can be shown (using both properties of the strong solution) that any other mechanism would cease to be incentive compatible when the other players infer that the principal is in the set of types that prefer it over the strong solution. Thus, although the strong solution is not necessarily the best incentive-compatible mechanism for any one type of the principal, it is the inscrutable compromise that all types must feel compelled to advocate.

For example, suppose that player 1, the principal, is selling a car which player 2 thinks may be worth either $1000 or $2000 to him, depending on its quality (bad or good) which player 1 already knows. In either case, suppose that the car is worth $0 to 1. Player 2 thinks that good or bad are equally likely, given his current information, but he can determine the quality with a costless inspection. Assume that 1 is risk-averse, but 2 is risk-neutral. The ex-ante optimal mechanism for 1 before he learned the quality, would have been to demand that 2 pay $1500 for the car without inspecting it, but it would be absurd for 1 to try to negotiate such a deal when he already knows the quality. In technical terms, the constant $1500 mechanism is not a strong solution, because 2 would not want to accept the deal if he knew that 1 had a bad car. The strong solution for 1 is, of course, to demand $1000 if bad and $2000 if good and let 2 check the quality of the car.

Any general theory of intertype compromise in negotiations should generalize the rule that a principal with private information advocates his strong solution, if one exists. Since the strong solution is not necessarily ex-ante efficient, we must not assume that cooperative solutions should be ex-ante efficient, just interim incentive-efficient.

5. Neutral bargaining solutions for games with incomplete information.

The derivation of Nash's [1950] bargaining solution, for two-person games with complete information, can be divided into two parts. First, the symmetry and efficiency axioms identify what the solutions must be for a class of relatively straightforward problems. Then these solutions are extended to all other problems by applying some independence axioms, which assert that certain ("irrelevant") transformations of the problem should not change the solution. The derivation of neutral bargaining solutions by Myerson [1983, 1984a] for games with incomplete information proceeds analogously, beginning with the analysis of a class of straightforward games that have strong solutions (axiom (1) below), and then extending the analysis to other games by independence axioms (axioms (2) and (3) below).

For two-person games with incomplete information, with a fixed disagreement point and actions that can be regulated, the neutral bargaining solution of Myerson [1984a] is defined as the smallest solution concept satisfying the following three properties. (1) If there is a strong solution for each player that leaves the other player at his disagreement payoff (so that it is obvious that each player would demand this strong solution if he were the principal) and if a .50-.50 randomization between these two strong solutions is interim incentive-efficient, then this .50-.50 random dictatorship (being equitable and efficient) should be a bargaining solution. (2) If μ is an interim incentive-efficient mechanism for a game Γ and if, for each $\varepsilon > 0$, there exists some game Γ' such that a bargaining solution for Γ' gives each type t_i of each player i a payoff not more than $U_i(\mu|t_i) + \varepsilon$, and Γ' differs from Γ only in that Γ' allows the players more jointly feasible actions than Γ, then μ must be a bargaining solution for Γ. (3) The set of solutions is invariant under decision-theoretically irrelevant transformations of the utility and probability functions. General existence of these neutral bargaining solutions can be shown.

The neutral bargaining solutions can be characterized in terms of a concept called virtual utility. It is well known that the problem of satisfying economic constraints can be decentralized efficiently if the constrained quantities are multiplied by some (carefully chosen) shadow prices and then added into the individuals' payoff functions. This idea can also be applied to the problem of satisfying incentive constraints. The resulting transformation of individuals' payoff functions is called virtual utility by Myerson [1984a, 1985a]. We may say that a type s_i jeopardizes another type t_i of player i if there is a positive shadow price assigned to the incentive constraint that type s_i of player i should not gain by pretending to be type t_i. Then, qualitatively, the virtual-utility payoffs for type t_i of player i differ from the real payoffs in a way that exaggerates the difference from the (false) types that jeopardize t_i. An incentive-compatible mechanism μ is incentive-efficient if and only if there exist virtual utility scales such that (i) only incentive constraints that bind in μ have positive shadow prices, and (ii) the players' sum of virtual utility is always maximized ex post in μ. That is, any incentive-efficient mechanism is virtually efficient ex post, for appropriately chosen virtual utility scales. If, in addition, each type of each player gets an expected virtual gain (over disagreement) which, in these virtual-utility scales, is equal to the expected virtual gain of the other player, then this mechanism is a neutral bargaining solution. That is, neutral bargaining solutions are incentive-efficient mechanisms that are also virtually equitable.

The concept of (weighted) utility transfers has played a crucial role in the analysis of multi-player cooperative games with complete information. For games with incomplete information, however, transfers may be more complicated phenomena because they can serve a signalling role as well as a redistributive role. (I may give someone a gift to prove something about myself, as well as to repay a debt.) Myerson [1984b] has shown that, in a technical sense, the natural extension of the concept of weighted-utility transfers to games with incomplete information is the concept of state-dependent virtual-utility transfers, because these concepts both parametrically generate families of linear activities that can transform the efficient frontier into a hyperplane that supports any arbitrarily chosen incentive-efficient mechanism of the original game without transfers. Using this insight, a natural extension of the Shapley NTU value to multi-player games with incomplete information has also been defined by Myerson [1984b] in terms of a coincidence of virtual efficiency and virtual equity, with respect to some virtual utility scales.

The neutral optima for a principal with private information are defined axiomatically to be a natural extension of the principal's strong solutions, and are shown to always exist by Myerson [1983]. To get some intuition about these neutral optima, recall that strong solutions were required to be safe and undominated for the principal. Since strong solutions often do not exist, existence requires that we relax at least one of these requirements. In the definition of neutral optima, the safeness requirement is relaxed. That is, the principal's neutral optima are defined so that they are, in some sense, as close to "safe" as possible within the set of incentive-compatible mechanisms that are undominated for the principal. For example, the principal's neutral optimum is the mechanism with the least amount of pooling among the principal's undominated mechanisms in a trading game studied by Myerson [1985b] (the uniform additive "bargaining for a lemon" game).

6. An example.

Suppose that a monopolistic seller is facing a monopsonistic buyer of a unique indivisible object, and the following facts are common knowledge. The object may be worth $0 or $80 to the seller, and the buyer assigns subjective probabilities .75 and .25 to these events, respectively. The object may be worth $20 or $100 to the buyer, and the seller assigns subjective probabilities .25 and .75 to these events, respectively. Each individual knows his own value for the object.

An ex-post efficient mechanism for this example would have the buyer promptly get the object except in the event that it is worth $80 to the seller and $20 to the buyer, which happens (from our outsider's perspective) only with probability 0.0625. However, it can be shown that there is no incentive-compatible mechanism that is ex-post efficient in this sense. That is, in any equilibrium of any trading game derived from this situation (by specifying in detail the rules by which terms of trade are proposed and ratified), there must be a positive probaility that the seller keeps the object even though it is worth more to the buyer.

To understand this impossibility result, let us consider a class of trading mechanisms that treat the buyer and seller symmetrically. (If there were any ex-post efficient incentive-compatible mechanism, there would be a symmetric one.) If the object is worth $0 to the sellɩr and $100 to the buyer then, for symmetry and ex-post efficiency, let trade occur for sure at price $50. If the object is worth $80 to the seller and $20 to the buyer, then let no trade occur. Now, let q denote the probability that trade occurs if the object is worth $0 to the seller and $20 to the buyer, and let $x denote the price of the object if trade occurs in this event. For symmetry, let q also be the probability that trade occurs if the object is worth \$80 to the seller and \$100 to the buyer, and let \$(100 - x) be the price if trade occurs in this event. To give each type of each trader an incentive to participate in the mechanism, we must have $x \leq 20$ (otherwise, a low-type buyer would have negative expected gains from trade). To give a low-type seller an incentive to report his type honestly, we must have

$$.75(50) + .25xq \geq .75(100 - x)q,$$

which implies that $q \leq 37.5/(75 - x)$. Thus, it is impossible to have $q = 1$, as ex-post efficiency would require.

The mechanisms in this class that satisfy

$$q = 37.5/(75 - x)$$

are all interim incentive-efficient, so these are the natural mechanisms to consider in searching for an equitable arbitrated or negotiated solution to this trading problem. The _strong_ types (that is, the seller's type with a reservation price of \$80, and the buyer's type with a reservation price of \$20) each prefer the mechanisms with lower x. The _weak_ types (the seller's type with a reservation price of \$0, and the buyer's type with a reservation price of \$100) prefer the mechanisms with higher x. Thus, the question of intertype compromise is essential to the analysis of this example.

In all of the symmetric incentive-efficient mechanisms described above, the weak type of a trader is indifferent between honestly reporting his type and claiming that he is strong, given his own information. But for each of these mechanisms, a weak-type trader would strictly prefer to pretend that he was strong if he knew that the other trader was weak. Thus, none of these mechanisms could be implemented by a process of unmediated bargaining in which the players, speaking one at a time and directly to each other, alternately announce offers until one accepts the other's latest offer. (Since the price in the accepted offer would have to depend on both players' types in any of these mechanisms, there would have to be some informative announcements before the end of bargaining. But, after the first such informative announcement, if it gave evidence of the announcer's weakness then the other trader could gain by always henceforth acting like a strong type.) Thus, as Farrell [1983] has observed, the use of a mediator may be essential to implement socially desirable mechanisms in a game. Confidential mediation is more than just a simplifying technical assumption; it is an important and practical way to let individuals reveal information or make concessions without reducing other individuals' incentives to also make concessions.

The ex-ante criterion would suggest that the symmetric mechanism with the highest possible x, $x = 20$ (and so $q = .68$) would be desirable, since this maximizes the ex-ante expected gains, before the traders learn their types. However, the traders already know their reservation prices, so this ex-ante criterion is not relevant to them.

The symmetric mechanism with $x = 0$ and $q = .5$ is the neutral bargaining solution for this example. That is, the neutral bargaining solution resolves the intertype-compromise question in favor of the strong types. This may seem reasonable in view of the fact that the problematic incentive constraints in this example are the constraints that a weak-type trader should not gain by pretending that he is strong. That is, in negotiations, we might expect to see a weak ($0) type of seller trying to pretend that he is really the strong ($80) type, but we would not expect to see a strong type pretending that he is weak. (Sellers may often overstate their reservation prices, but they usually do not understate them!) Thus it is reasonable that the traders should negotiate to the symmetric mechanism that is most preferred by the strong types. Since only the weak types feel the need to be inscrutable, the inscrutable compromise gets resolved completely in favor of the strong types. (If negative prices were allowed, the strong types would actually prefer to implement a symmetric mechanism with $x < 0$. But a strong trader would need to be inscrutable to negotiate for such strange mechanisms with negative x, since he would be offering negative gains to the other trader. Thus, the inscrutable compromise would still be at $x = 0$.)

All of the interim incentive-efficient symetric mechanisms described above (with $q = 37.5/(75 - x)$) are optimal solutions to the mathematical problem of maximizing

$$(11/16)(U_s(\mu|0) + U_b(\mu|100)) + (5/16)(U_s(\mu|80) + U_b(\mu|20))$$

subject to the constraint that μ should be an incentive-compatible mechanism. (Here, $U_s(\mu|\theta)$ denotes the seller's expected utility from the mechanism μ if his value for the object is $\$\theta$; similarly, $U_b(\mu|\theta)$ denotes the buyer's expected utility from μ if his value is $\$\theta$. Notice that the objective-function coefficients have been chosen so that the weights given to the strong types are slightly larger than the ex ante probabilities of these types.) In this optimization problem, the weak types jeopardize the strong types, and the shadow prices of these incentive constraints is 1/16. With these coefficients and shadow prices, the virtual utility of each weak type is the same as his actual utility, but the virtual utilities of the strong types exaggerate their difference from the weak types. Specifically, the strong type of the seller has a virtual value of $100 (greater than his actual value of $80) for the object, and the strong type of buyer has a virtual value of $0 (less than his actual value of $20) for the object. With these virtual values, there are no virtual gains from trade when one trader is weak and the other trader is strong, so the mechanisms that randomly determine whether trade occurs or not in such cases are virtually efficient ex post, as incentive-efficiency requires. But only the symmetric incentive-efficient mechanism with $x = 0$ and $q = .5$ is virtually equitable, since all the others have the strong type of buyer paying more (and the strong type of seller getting less) than his virtual value when trade occurs. So this mechanism satisfies the virtual-equity conditions for a neutral bargaining solution.

Myerson [1985b] noted that neutral bargaining solutions tend to have the following property, called arrogance of strength: if two individuals have symmetric bargaining ability but one individual's actual type is surprisingly strong, then the outcome of the neutral bargaining solution tends to be similar to what would have been the outcome if this strong individual had been the principal, except that the probability of the disagreement outcome is larger. That is, we may predict that a low-probability strong type will try to dictate his optimal agreement, and he will either be successful in this arrogant Boulware strategy or else he will fail and cause negotiations to break down. In this example, if the seller were the principal, he would (in either type) demand a price of $100, and the weak-type buyer would aquiesce to this price. The strong-type seller demands this same $100 price in the neutral bargaining solution, but the weak-type buyer acquiesces to it only with probability 1/2. (In the other symmetric mechanisms, the behavior of the strong types goes from "arrogant" to "timid" as x increases from 0 to 20.)

Using risk-neutrality, one can check that the neutral bargaining solution for this example is equivalent to a mechanism in which either the buyer or seller, each with probability 1/2, gets to make a first-and-final price offer.

Thus, the neutral bargaining solution for this example can be interpreted directly in terms of the random-dictatorship axiom of Myerson [1984a].

In this example, we have remarked that there is no incentive-compatible trading mechanism that guarantees that trade will promptly occur when the object is worth more to the buyer, and that trade will never occur when the object is worth more to the seller. Thus, we must recommend mechanisms that allow a positive probability of no trade (or, in a dynamic model, a costly delay before trading) when one trader is strong and the other is weak, even though there actually is a range of mutually acceptable prices. Let us now consider what would happen if a mediator ignored this prescription, and instead followed a policy of recommending trade at a mutually acceptable price, whenever he can identify one. For example, a mediator might ask each trader to report his value for the object (insisting that the seller must report $0 or $80 and the buyer must report $20 or $100, since these ranges are common knowledge) and then recommend trade at a price that is halfway between the two reported values whenever the buyer's reported value is higher. For simplicity, let us suppose that this mediator also has the power to regulate the price, so that his price recommendations must be followed if trade occurs, although he cannot force them to trade. Such a mediation plan induces a game that has three equilibria. In one equilibrium, the buyer always reports $20, even if he is weak, and the seller is honest, so that trade occurs at a price of $10 with probability 0.75. In another equilibrium, the seller always reports $80, even if he is weak, and the buyer is honest, so that trade occurs at a price of $90 with probability 0.75. In the third equilibrium, the strong type of each player is always honest, but the weak type randomizes between honestly reporting his value, with probability 20/75, and reporting the strong type's value, with probability 55/75. This third equilibrium treats the two traders symmetrically, unlike the other two equilibria, but the probability of trade in it is only 0.36. Thus, a mediation plan which fails to recognize binding incentive constraints can generate a communication equilibrium which is either very inequitable (like the first two equilibria here) or very inefficient (like the third). As a matter of practical advice, it is better for a mediator to plan to pay the cost of incentive compatibility in a controlled incentive-efficient way than to try to avoid paying the costs of incentive compatibility at all.

7. Dynamic models of bargaining.

The importance of ex-post efficiency in economic analysis is derived not only from welfare considerations but from stability considerations. That is, one might wonder how the mechanisms discussed in the preceding section could actually be implemented if the seller can keep making offers to sell the object. After all, if the seller still has the object and knows that there is a positive probability that the buyer might be willing to pay more than the seller's reservation price, why should the seller not make another (slightly lower) offer to sell? It seems that it may make a difference whether this is a one-stage game or a multi-stage game in which the players can continue to propose potential terms of trade.

To analyse a simple dynamic bargaining game in continuous time, suppose that, in our example, the traders will trade as soon as one of them concedes that he is weak, at a price of x dollars if the seller concedes first, and at a price of 100 - x dollars if the buyer concedes first (and at a price of

$50 if they both concede simultaneously). Here, we let x be some given constant between 0 and 20. Suppose that the traders have the same discount rate δ. While he retains the object, the seller gets a constant stream of benefits, which, if extended forever, would have a present discounted value equal to his value for the object. In this game, there is a symmetric equilibrium in which the strong types never concede, and the weak type of each player plans independently to concede somewhere between time 0 and time $(100 - 2x)(\ell n(4))/(\delta x)$, using the cumulative distribution function

$$F(\tau) = \min\{1,\ (4/3)(1 - e^{-\tau\delta x/(100-2x)})\}$$

to determine his time τ of concession. With this distribution, if one trader is weak and the other trader is strong, the expected time of concession by the weak trader is $(200 - 4x)/(3\delta x)$; letting $x = 10$ and $\delta = .10$ (when time is measured in years), this is 53.3 years! These long delays bfore trading greatly reduce the expected present discounted values of the gains for each type of each trader. In fact, for each x, the symmetric equilibrium of this dynamic bargaining game is Pareto-dominated (using the interim welfare criteria) by the corresponding symmetric mechanism (for the same x) that was discussed in the preceding section. The problem is that having opportunities to trade later reduces the incentive for players to make concessions now.

In general, a player's inalienable right to take various actions in the future (i.e., to quit a job, or to offer to sell something he owns) generates strategic incentive constraints which reduce the set of incentive-compatible mechanisms relative to what it would have been if, at the beginning of the game, the player had the option to precommit himself to never take advantage of these opportunities. The basic principles for analyzing such dynamic incentive constraints in multi-stage games are discussed by Myerson [1984c].

The key idea is that it suffices to consider centralized communication mechanisms in which, at every stage, each player confidentially reports his new information to a central mediator, who in turn confidentially recommends the action each player should take at the current stage, such that all players have an incentive to be honest and obedient to the mediator. Furthermore, sequential rationality is actually easier to characterize in games with communication than in games without communication (as in Kreps and Wilson [1982]), since Myerson [1984c] showed that a communication equilibrium is sequentially rational if and only if it avoids certain codominated actions. In a game where players move one at a time, these codominated actions can be identified simply by sequentially eliminating all dominated actions for all types of all players, beginning in the last stage and working backwards.

Rubinstein [1982] and others have shown that many natural bargaining games have a unique perfect equilibrium. Most commonly, the games in which this occurs are games in which the players have no private information and never make simultaneous moves. In the above terminology, such results assert that the dynamic incentive constraints reduce the set of incentive-compatible mechanisms to a single point, which is certainly a very important result whenever it holds. However, games with incomplete information and games in which players can make simultaneous moves generally have large sets of equilibria, even when sequential rationality is required. (In our example, it was rather unnatural to suppose that the price x was exogeonously given and constant over time. Once this assumption is dropped, the number of eqilibria increases enormously.) Such multiplicity is actually good news to an arbitrator (or a principal, or a set of negotiating players) who can determine the equilibrium to be played, since more equilibria means a larger set to select from. Our problem is to predict how impartial arbitration or effective negotiation might select among these equilibria.

Multi-stage dynamics do more than just generate incentive constraints, however. They also generate new welfare criteria that may influence the cooperative determination of which equilibrium or incentive-compatible mechanism is to be implemented. For example, an equilibrium that gives high expected payoffs to all players, from their perspective at the beginning of the game, might actually give them all low payoffs after some event which has positive probability. (See Abreu, Pearce, and Stacchetti [1984], for some examples.) In such cases, even though the equilibrium is sequentially rational for each individual, we might suspect that the players would try to jointly renegotiate to some other equilibrium that seems better after this event, so that the initial equilibrium plan should not have been credible in the first place. Similarly, if there is one principal player, who has a monopoly on the power to negotiate and can thus designate the equilibrium that everyone will play, he may find that his exercise of this power at the first stage of the game is constrained by the other players' perception that he would reexercise this power at a later stage, after some event that makes his former selection seem less attractive to him than some other equilibrium. To understand such issues, we will need to develop a theory of recursively negotiated games, in which (some or all of) the players at each stage can negotiate their current and future behavior, subject to the usual incentive constraints plus the constraints implied by renegotiation opportunities in later stages.

REFERENCES

D. Abreu, D. Pearce and E. Stacchetti [1984], "Optimal Cartel Equilibria with Imperfect Monitoring," discussion paper, Harvard University.

K. Binmore [1981], "Nash Bargaining Theory II," discussion paper, London School of Economics.

V. P. Crawford [1985], "Efficient and Durable Decision Rules,: a Reformulation," Econometrica 53, 817-835.

J. Farrell [1983], "Communication in Games I: Mechanism Design without a Mediator," discussion paper, Massachusetts Institute of Technology.

R. Fisher and W. Ury [1981], Getting to Yes, Boston: Houghton Mifflin.

J. C. Harsanyi [1967-8], "Games with Incomplete Information Played by 'Bayesian' Players," Management Science 14, 159-182, 320-334, 486-502.

J. C. Harsanyi [1975], "The Tracing Procedure: a Bayesian Approach to Defining a Solution for n-Person Games, International Journal of Game Theory 4, 61-94.

B. Holmström and R. B. Myerson [1983], "Efficient and Durable Decision Rules with Incomplete Information," Econometrica 51, 1799-1819.

E. Kohlberg and J. F. Mertens [1983], "On the Strategic Stability of Equilibria," CORE Discussion Paper No. 8248, Université Catholique de Louvain.

D. M. Kreps and R. Wilson [1982], "Sequential Equilibria," Econometrica 50, 863-894.

R. D. Luce and H. Raiffa [1957], Games and Decisions, New York: Wiley.

R. B. Myerson [1982], "Optimal Coordination Mechansims in Generalized Principal-Agent Problems," Journal of Mathematical Economics 11, 67-81.

R. B. Myerson [1983], "Mechanism Design by an Informed Principal," Econometrica 51, 1767-1797.

R. B. Myerson [1984a], "Two-Person Bargaining Problems with Incomplete Information," Econometrica 52, 461-487.

R. B. Myerson [1984b], "Cooperative Games with Incomplete Information," International Journal of Game Theory 13, 69-96.

R. B. Myerson [1984c], "Multistage Games with Communication," Discussion Paper No. 590, Northwestern University, to appear in Econometrica.

R. B. Myerson [1984d], "An Introduction to Game Theory," Discussion Paper No. 623, Northwestern University.

R. B. Myerson [1985a], "Bayesian Equilibrium and Incentive Compatibility: An Introduction," discussion paper, Northwestern University, to appear in Social Goals and Social Organization, edited by L. Hurwicz, D. Schmeidler, and H. Sonnenschein, Cambridge University Press.

R. B. Myerson [1985b], "Analysis of Two-Person Bargaining Games with Incomplete Information," discussion paper, Northwestern University, to appear in Game Theoretic Models of Bargaining, edited by A. E. Roth, Cambridge University Press.

J. F. Nash [1950], "The Bargaining Problem," Econometrica 18, 155-162.

J. F. Nash [1951], "Noncooperative Games," Annals of Mathematics 54, 289-295.

J. F. Nash [1953], "Two-Person Cooperative Games," Econometrica 21, 128-40.

A. Rubinstein [1982], "Perfect Equilibrium in a Bargaining Model," Econometrica 50, 97-109.

R. Selten [1975], "Reexamination of ther Perfectness Concept for Equilibrium Points in Extensive Games," International Journal of Game Theory 4, 25-55.

T. C. Shelling [1960], The Strategy of Conflict, Harvard University Press.