Myerson, Roger B.

Working Paper

An Introduction to Game Theory

Discussion Paper, No. 623

Provided in Cooperation with:
Kellogg School of Management - Center for Mathematical Studies in Economics and Management Science, Northwestern University

Suggested Citation: Myerson, Roger B. (1984) : An Introduction to Game Theory, Discussion Paper, No. 623, Northwestern University, Kellogg School of Management, Center for Mathematical Studies in Economics and Management Science, Evanston, IL

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

Discussion Paper No. 623

AN INTRODUCTION TO GAME THEORY

by

Roger B. Myerson

September 1984
revised July 1985

Abstract: Game theory is the study of mathematical models of conflict and cooperation between intelligent rational decision makers. This paper is an introduction to game theory for economists and other social scientists. No attempt is made to survey all major results in the literature. Instead, the goal is to present a unified development of the important fundamental ideas of game theory. Topics include: decision-theoretic foundations of game theory, basic models of games, Nash equilibria, extensions and refinements of the equilibrium concept, the Nash bargaining solution, cooperative games with and without transferable utility, and cooperative games with incomplete information.

Acknowledgements: The author is deeply indebted to Ehud Kalai and Robert Weber for many long discussions about game theory. Support by the John Simon Guggenheim Memorial Foundation and the Alfred P. Sloan Foundation is gratefully acknowledged.

To appear in Studies in Mathematical Economics, edited by Stanley Reiter, to be published by the Mathematics Association of America.

AN INTRODUCTION TO GAME THEORY

by

Roger B. Myerson

1. The Decision-Theoretic Foundations of Game Theory

To understand the fundamental ideas of game theory, one should begin with a review of decision theory. Decision theory is concerned with the problem of one individual who has to choose among various risky options, which may be called "lotteries." Each lottery would give the individual a randomly determined outcome or "prize," possibly depending upon some unknown factors which we may call the "state" (or "state of the world"). Using remarkably weak assumptions about how a rational decision maker should behave, it has been shown that such a decision maker should be able to assess subjective probability numbers $p(s)$ for every possible state s, and utility numbers $u(x)$ for every possible prize x, such that he always prefers to choose the lottery that has the highest subjective expected utility. The subjective expected utility of a lottery is defined by the formula

$$\sum_x \sum_s p(s)\ f(x|s)\ u(x),$$

where $f(x|s)$ denotes the objective probability that the lottery would give prize x if s were the true state of the world. (Classic and seminal presentations of this result are in von Neumann and Morgenstern [1944], Savage [1954], and Raiffa [1968]. See also Luce and Raiffa [1957, chapter 2] and Myerson [1979].)

This result assures us that the behavior of a rational decision maker can

be described mathematically, for both theoretical and practical purposes, if these probability and utility functions can be assessed. However, suppose that one of the factors that is unknown to some given individual (#1) is the action of some other individual (#2). To assess a subjective probability distribution over 2's possible actions, individual 1 may try to imagine himself in individual 2's position. But in this thought experiment, he may realize that 2 is trying to solve a rational decision-making problem of his own, and that problem involves assessing a subjective probability distribution over 1's possible actions. Thus, the rational solution to each individual's decision problem depends on the solution to the other individual's problem, and neither can be solved without the other. So when rational individuals interact, their decision problems must be analyzed together, like a system of equations. Such analysis is the subject of game theory.

Game theory can be defined as the study of mathematical models of conflict and cooperation between intelligent rational decision makers. By "rational" we mean that each individual's decision-making behavior would be consistent with the maximization of subjective expected utility, if the other individuals' decisions were specified. By "intelligent" we mean that each individual understands everything about the structure of the situation that we theorists understand, including the fact that all other individuals are intelligent rational decision makers. Thus, if we develop a theory that describes how the players in some game should behave, then we must assume that each player in the game will also understand this theory and its predictions.

It may be useful to compare game theory to price theory, for example. In the general equilibrium model of price theory, it is assumed that every individual is a rational utility-maximizing decision maker, but it is not assumed that individuals undertand the whole structure of the economic model

that the price theorist is studying. In price-theoretic models, individuals only perceive and respond to some intermediating market signals. In game theory, we assume that all individuals perceive and respond to each other directly. Thus, game theory may be better than price theory for describing markets with relatively few participants, as in oligopolistic competition or union/management relations. On the other hand, game theory is generally worse than price theory for describing the macroeconomy, in which even the assumption that individuals know all the prices may be too strong.

Of course, the game theorist's assumption that all individuals are perfectly rational and intelligent (in the above sense) may never be satisfied in any real life situation. But on the other hand, we should be suspicious of theories and predictions that are not consistent with this assumption. That is, if a theory predicts that some individuals will be systematically fooled or led into mistakes that hurt themselves, then this theory will tend to lose its validity when these individuals learn to better understand the situation. The importance of game theory in the social sciences is derived from this fact.

2. Basic Models of Game Theory

The most general models used to describe games are dynamic models, which describe all the sequences of actions or moves that could be made by the players over time during the play of the game. Kuhn [1953] developed the formal definition of the extensive form, which is now the standard dynamic model in the literature on game theory. (See Luce and Raiffa [1957, chapter 3], Owen [1982, chapter 1], Shubik [1982, chapter 3], and Kreps and Wilson [1982].) For our purposes here, however, it will suffice to discuss a somewhat simpler multistage form (used in Myerson [1984d]).

To describe a game in multistage form, we must first specify the set of sequentially numbered stages ($\{1,2,\ldots,K\}$) and the set of players in the game. We let $N = \{1,2,\ldots,n\}$ denote the set of players, with i denoting a member of N. For each stage k and for each player i, we must specify the set of possible signals (or new information) that player i could get at the beginning of stage k, and the set of possible actions (or moves) that player i could choose at the end of stage k. An information state for player i in stage k is any possible sequence of signals that he might have gotten in the first k stages and of actions that he might have taken in the first k - 1 stages. If we assume that each player has perfect recall, then such an information state would characterize what player i knows at the beginning of stage k. For any stage k before the last, and for any possible combination of the players' information states and actions at stage k, we must specify the probability of each possible combination of new signals for the players at the beginning of the next stage k + 1. For stage 1, we must specify the probability of each possible combination of signals for the players at the beginning of the first stage. The set of outcomes of the game is the set of all possible sequences of signals and actions for all players at all the stages of the game. For every player, we must specify a payoff function, which assigns a utility value to each outcome of the game. These payoff functions describe the players' preferences, and complete our specification of the multistage game.

A game in strategic form is a special case of the multistage form in which there is only one stage and each player has only one possible information state. That is, to define a game in strategic form, we need to specify a set of players ($N = \{1,2,\ldots,n\}$), and, for each player i, we must specify a set of possible actions or strategies (C_i) and a payoff function (u_i). Here, each player's payoff function is a map from the set of possible

combinations of actions for all the players ($C_1 \times \ldots \times C_n$) into the set of real numbers. That is, $u_i(c_1,\ldots,c_n)$ denotes the utility value, for player i, of the outcome of the game when $(c_1,\ldots,c_n)$ is the combination of players' actions (each player j using c_j).

Von Neumann and Morgenstern [1944] argued that there may be no loss of generality in restricting our theoretical attention to these conceptually simpler games in strategic form. Given any multistage game, they showed how to construct an equivalent game in strategic form (which is also called the normal form of the multistage game). A strategy for a player in a multistage game is any function that specifies a feasible action for the player, at every stage and every possible information state. That is, a strategy for player i is a complete plan of action for player i, at all stages and all possible information states in the multistage game. The set of actions for any player in the equivalent strategic-form game is defined to be his set of strategies in the given multistage game. For any combination of strategies $(s_1,\ldots,s_n)$, the payoff $u_i(s_1,\ldots,s_n)$ to any player i in the strategic-form game is defined to be his expected payoff in the multistage game when all players plan to use their given strategies (each player j using his strategy s_j).

Thus, when we reduce a multistage game to strategic form, we suppress its dynamic structure and condense all decision making into one stage. This is a major simplification in the conceptual structure of our model. However, the set of strategies is sometimes so large that it may be more practical to study the dynamic model than the strategic form. (For example, the set of all possible strategies for each player in chess is a finite but astronomically large set.)

When we theoretically analyze a multistage game, we are trying, before the game begins, to predict what each player should do at each stage and each

possible information state. If each player is as intelligent as we are, then he should also be able, before the game, to analyze the game and rationally plan all of his actions in all possible events. But if all players choose their strategies in advance (at stage "0") then the equivalent strategic form is a precise description of their decision problems. That is, when players plan their actions in advance, they are taking their decision-making process outside of the dynamic structure of the game. This strongly suggests that there may be no theoretical loss in reducing multistage games to the conceptually simpler strategic form.

This insight is very important in game theory, even though its limitations are now becoming re-examined in the literature (Selten [1975], Kreps and Wilson [1982], Kohlberg and Mertens [1983], and Myerson [1984d]). For example, sometimes a theorist may try to defend a general solution concept for strategic-form games by arguing that players would converge to it in a game that has been repeated many times. But such an argument would ignore the fact that a repeated game is just a kind of multistage game, and so it can be reduced to one large strategic-form game itself. If the general solution concept is valid then it should be applied to this overall game, not to the repeated stages separately.

The argument for reducing a game to strategic form relied on the assumption that each player could plan his actions before getting any private information or signals. However, some parts of a player's private information may be so basic to his identity that it is not meaningful to talk about him planning his actions before learning this information (e.g., what is the player's gender, native language, and level of risk aversion). Harsanyi [1967-8] called such initial information the <u>type</u> of a player. If the players have any uncertainty about each others' types, then it may not be possible to

completely reduce a game to strategic form. Instead, we must use a somewhat more general class of models, which Harsanyi called Bayesian games.

To define a Bayesian game, we must specify a set of players ($N = \{1,2,\ldots n\}$), and for each player i we must specify a set of possible actions (C_i), a set of possible types or information states (T_i), a probability function (p_i), and a payoff function (u_i). For every possible type t_i of player i, the probability function p_i must specify a probability distribution $p_i(\cdot|t_i)$ over the set of all possible combinations of other players' types ($T_{-i} = T_1 \times \ldots \times T_{i-1} \times T_{i+1} \times \ldots \times T_n$), which represents what player i would believe about the other players' types if his type were t_i. That is, $p_i(t_1,\ldots,t_{i-1},t_{i+1},\ldots,t_n|t_i)$ (or $p_i(t_{-i}|t_i)$) denotes the subjective probability that i would assign to the event that player 1's type is t_1, 2's type is t_2, etc., when i knows that his own type is t_i. The payoff function for player i must specify a numerical utility value $u_i(c_1,\ldots,c_n,t_1,\ldots,t_n)$ to every possible combination of players' actions $(c_1,\ldots,c_n)$ and every possible combination of players' types $(t_1,\ldots,t_n)$.

When we study a Bayesian game, we assume that the sets and functions specified above are common knowledge among the players (that is, every player knows these structures, every player knows that every player knows them, etc.; see Aumann [1976]). In addition, each player knows his own actual type. Thus, the type of player i is defined to be a variable that represents all of his initial information (about preferences and endowments and other players' beliefs) that may be unknown to other players and to us theorists. Mertens and Zamir [1983] showed that, in principle, it should always be possible to mathematically construct type-sets that are large enough to subsume all possible states of each player's initial private information, so that the Bayesian game model has complete theoretical generality among nondynamic

models. In some applications, of course, these type sets may be too large for tractable analysis, which may limit the practical applicability of the Bayesian game model. On the other hand, if each player has only one possible type, then the Bayesian game reduces to a game in strategic form. If there is at least one player who has more than one possible type, then the game is said to be a game with incomplete information.

Harsanyi [1967-8], following a suggestion by R. Selten, discussed a formal way of reducing any Bayesian game to a game in strategic form, using what he called the Selten model. In the Selten model, there is one player for every possible type of every player in the original Bayesian game. Thus, for a Bayesian game with three players, each of whom has five possible types, there are fifteen players in the Selten model. The set of actions for each player in the Selten model is the same as the set of possible actions for that player in the original game for whom the Selten-model player represents one possible type. The payoffs to players in the Selten model are defined to be the expected payoffs for the corresponding types of players in the original Bayesian game. That is, if we let $\gamma_j(t_j)$ denote the action selected by the Selten-model player for each type t_j of each player j in the Bayesian game, then the corresponding payoff to the Selten-model player for any type t_i of any i would be

$$\sum_{t_{-i}} p_i(t_{-i}|t_i)\, u_i(\gamma_1(t_1),\ldots,\gamma_n(t_n),t_1,\ldots,t_n).$$

For a more detailed introduction to the analysis of Bayesian games, see Myerson [1985a].

3. Nash Equilibria

Nash's [1951] definition of equilibrium is probably the most important concept in game theory. Given a strategic-form game, a combination of actions or strategies for the players is a Nash equilibrium if each player's action maximizes his expected utility, given the actions of the other players.

To prove existence of equilibria, it is necessary to allow players to randomize. A randomized strategy for player i is any probability distribution over the set of possible actions for player i. If σ_i is a randomized strategy for player i, then $\sigma_i(c_i)$ denotes the probability that player i will select the action c_i in the randomized strategy σ_i. The players are assumed to randomize independently in a game without communication, so that player i's expected payoff would be

$$u_i(\sigma_1,\ldots,\sigma_n) = \sum_{(c_1,\ldots,c_n)} \left(\prod_{j=1}^{n} \sigma_j(c_j)\right) u_i(c_1,\ldots,c_n),$$

when the players used the randomized strategies $(\sigma_1,\ldots,\sigma_n)$. (Here Π represents the multiplicative product.) A combination of randomized strategies $(\sigma_1,\ldots,\sigma_n)$ is a (Nash) equilibrium if, for every player i and every randomized strategy $\hat{\sigma}_i$,

$$u_i(\sigma_1,\ldots,\sigma_n) \geqslant u_i(\sigma_1,\ldots,\sigma_{i-1},\hat{\sigma}_i,\sigma_{i+1},\ldots,\sigma_n).$$

That is, each player i cannot increase his expected payoff by using any other randomized strategy $\hat{\sigma}_i$ instead of σ_i, when every other player j is using σ_j. Nash proved that, for any strategic-form game in which the set of players and the sets of possible actions are all finite, there exists at least one equilibrium in randomized strategies. It is appropriate to call this existence theorem the fundamental theorem of game theory.

For a Bayesian game with incomplete information, Harsanyi [1967-8] defined Bayesian equilibria to be the Nash equilibria of the equivalent Selten-model game in strategic form. That is, a Bayesian equilibrium specifies an action or randomized strategy for each type of each player, so that each type of each player would be maximizing his own expected utility, over all his possible actions, when he knows his own type but does not know the other players' types. Notice that, in a Bayesian equilibrium, a player's action can depend on his own type, but not on the types of the other players. (By definition, a player's type is supposed to subsume all of his private information at the beginning of the game, when he chooses his action or strategy.) We need to specify what every type of every player would do, not just the actual types, because otherwise we could not define the expected payoff for a player who does not know the other players' actual types.

The importance of Nash (and Bayesian) equilibria comes from the following argument. Suppose that we are acting either as theorists, trying to predict the players' behavior in a given game, or as social planners, trying to prescribe the players' behavior. If we specify what strategies should be used by the players, and if the players understand this specification also (recall that they know everything that we know about the game), then we must either specify an equilibrium or impute irrational behavior to some players. If we are not specifying an equilibrium, then some player could gain by changing his strategy. Thus, a non-equilibrium specification would be a self-denying prophecy if the players all believed it.

This argument uses the assumption that the players in a strategic form game are choosing their actions or strategies independently, so that one player's change of strategy cannot cause a change by any other player. In a sense, this independence assumption is without loss of generality. If there

are rounds of communication between the players, then the set of strategies for each player in the strategic-form game can be redefined to include all plans for what to say in these rounds of communication and what actions to choose, depending on the previously received messages. That is, a game with pre-play communication can be viewed as an extensive or multistage game, and can be reduced to an equivalent strategic-form game as described in Section 2. (On the other hand, we will see that it is often more convenient to omit such possibilities for communication from the structure of the game and to build them into the solution concept instead, which will take us to the concept of correlated equilibrium in Section 5.)

Aumann and Maschler [1972] reexamined the argument for Nash equilibrium as a solution for games like the following example (in strategic form):

		Player 2	
		L	R
Player 1	T	0,0	0,-1
	B	1,0	-1,3

Example 1

(Here $C_1 = \{T,B\}$, $C_2 = \{L,R\}$ and the numbers in each box are the utility payoffs (u_1,u_2).) The unique equilibrium for this game is the pair of randomized strategies (σ_1,σ_2), where $\sigma_1(T) = 3/4$, $\sigma_1(B) = 1/4$, $\sigma_2(L) = 1/2$ and $\sigma_2(R) = 1/2$. However, Aumann and Maschler suggest that player 1 might prefer to choose T and player 2 might prefer to choose L (each with probability one), because these actions are optimal responses to the equilibrium strategies and guarantee each player his expected equilibrium payoff of zero. But if such

behavior were correctly anticipated then player 1 would be irrational not to choose B, because it is his unique best response to L. Thus, a theory that predicts the actions T and L in this game would destroy its own validity, because (T,L) is not a Nash equilibrium.

Notice that we have not given any direct argument as to why intelligent rational players must use equilibrium strategies in a game. When someone asks why players in a game should behave as in some Nash equilibrium, this author's favorite response is to ask "why not?" and to let the challenger specify what he thinks the players should do. If this specification is not a Nash equilibrium, then (as above) we can show that it would destroy its own validity if the players believed it to be an accurate description of each others' behavior. It may be better to think of Nash equilibrium as a "pre-solution concept," rather than as a solution concept, because being a Nash equilibrium is only a necessary condition, not a sufficient condition, for being a good prediction of rational players' behavior. That is, every outcome that is not an equilibrium will necessarily be an unreasonable prediction of how intelligent rational decision makers would behave. Thus, the concept of Nash equilibrium imposes a contraint on social planners and theorists, in that they cannot predict nonequilibrium behavior.

Equilibria in randomized strategies sometimes seem difficult to interpret. It is easy to check (by examining the four possibilities) that there is no equilbrium without randomization in Example 1. But the necessity for player 1 to randomly choose among T and B with probabilities 3/4 and 1/4, respectively, might not seem to coincide with any compulsion that people experience in real life. Of course, if player 1 thinks that player 2 is equally likely to choose L or R then player 1 is willing to randomize in any way between T and B. But what could make player 1 actually want to use the

precise probabilities 3/4 and 1/4?

Harsanyi [1973] showed that Nash equilibria that involve randomized strategies can be interpreted as limits of equilibria in which each player is (almost) always choosing his a uniquely optimal action. Harsanyi's basic idea is to modify the game so that each player has slightly different information about the payoffs. (See also Milgrom and Weber [1984].) For example, suppose that Example 1 were modified slightly, to the following game with incomplete information:

		Player 2	
		L	R
Player 1	T	$\varepsilon\tilde{\alpha},\varepsilon\tilde{\beta}$	$\varepsilon\tilde{\alpha},-1$
	B	$1,\varepsilon\tilde{\beta}$	$-1,3$

Example 1a

Here $0 < \varepsilon < 1$, and $\tilde{\alpha}$ and $\tilde{\beta}$ are independent and identically distributed, each with a uniform distribution over the interval from 0 to 1. When the game is played, player 1 knows the value of $\tilde{\alpha}$ but not $\tilde{\beta}$, and player 2 knows the value of $\tilde{\beta}$ but not $\tilde{\alpha}$. If ε is zero then Example 1a becomes exactly the same as Example 1, so let us think of ε as a very small positive number (say, 10^{-9}). Then $\tilde{\alpha}$ and $\tilde{\beta}$ can be interpreted as minor factors that have a very small influence on the players' payoffs when T or L is chosen.

Given ε, there is a unique Bayesian equilibrium for Example 1a. Player 1 chooses T if he observes $\tilde{\alpha}$ greater than $(2 + \varepsilon)/(8 + \varepsilon^2)$, and he chooses B otherwise. Player 2 chooses L if he observes $\tilde{\beta}$ greater than $(4 - \varepsilon)/(8 + \varepsilon^2)$, and he chooses R otherwise. In these equilibrium strategies, each player always gets strictly higher utility from the action

that he is choosing than he would get from the other action (except in the zero-probability event that $\tilde{\alpha} = (2 + \varepsilon)/(8 + \varepsilon^2)$ or $\tilde{\beta} = (4 - \varepsilon)/(8 + \varepsilon^2)$). That is, each player's expected behavior makes the other player almost indifferent between his two actions, so that the minor factor that he observes independently can determine a unique optimal action for him. Notice that, as ε goes to zero, this equilibrium converges to the unique equilibrium of Example 1, in which player 1 chooses T with probability 3/4 and player 2 chooses L with probability 1/2.

Thus, in general, when we study an equilibrium involving randomized strategies, we may interpret each player's randomization as depending on minor factors that have been omitted from the description of the game. Or, to put it another way, when a game has no nonrandom equilibria, we should expect that a player's optimal action may be determined by some minor factors that he observes independently of the other players.

Two general observations about Nash equilibria are now in order. Nash equilibria may be <u>nonunique</u>; and Nash equilibria may be <u>inefficient</u>.

For an example of inefficiency, consider the following games, known as the "Prisoner's Dilemma":

		Player 2	
		L	R
Player 1	T	5,5	0,6
	B	6,0	1,1

Example 2

In this game, (B,R) is the unique Nash equilibrium, but it is also the only

outcome of the game that is not Pareto efficient. (See Luce and Raiffa [1957] for the story behind the names of this and the next example.)

For an example of nonuniqueness, consider the following game, known as the "Battle of the Sexes":

		Player 2	
		L	R
Player 1	T	2,1	0,0
	B	0,0	1,2

Example 3

There are three equilibria of this game: (T,L), which player 1 prefers; (B,R), which player 2 prefers; and a randomized equilibrium (2/3[T] + 1/3[B], 1/3[L] + 2/3[R]), which gives each player an expected utility of 2/3. The third equilibrium is also an example of inefficiency, since both players would prefer (T,L) or (B,R).

For games with multiple equilibria, anything (in the structure of the game or in the commonly observed environment in which it is played) that focuses the players' attentions on one particular equilibrium may create a situation in which all players expect this equilibrium and thus actually implement it. Schelling [1960] called this the <u>focal-point effect</u>. For example, if the players learned the Battle of the Sexes game from a book in which the payoff "1,2" was printed in red ink, all else being in black ink, then the (B,R) equilibrium would be much more likely to be played. Alternatively, suppose that player 1 is a woman and player 2 is a man in this game, and suppose that the players come from a culture in which women have

traditionally deferred to men. Then, even though this cultural tradition has no binding force on the players, it probably will cause both players to have the self-verifying expectation that player 2 will choose R and player 1 will choose B.

Another way that the players could become focused on one equilibrium, and so induced to implement it, is if some authoritative individual suggests it. If one player in the game has such authority or power of suggestion, so that he can select the equilibrium that all players will implement, then he may be called the principal of the game. (This definition is consistent with the usage of the term in most of the literature on principal-agent analysis.) An arbitrator is an outside individual, different from the players, who has power of suggestion to select the equilibrium.

4. Refinements of the Nash Equilibrium Concept

In some games with multiple equilibria, there may be some equilibria that seem intrinsically unreasonable for the players to implement, even if a principal or arbitrator tried to persuade the players to do so. For example, consider the following game:

		Player 2	
		L	R
Player 1	T	1,9	1,9
	B	0,0	2,1

Example 4

There are two equilibria (T,L) and (B,R); but it may be unreasonable to expect the players to implement (T,L). Player 2 can only gain by choosing R instead of L. (Remember that the players choose independently, so player 2 should not expected a switch from L to R to affect player 1's choice.) Then, if player 1 expects player 2 to choose R, B must be player 1's rational choice.

To identify and eliminate such unreasonable equilibria from our solution concepts, many refinements of the Nash equilibrium concept have been proposed. Three general criteria have been used to develop such refinements: (i) elimination of unreasonable actions; (ii) sequential rationality; and, (iii) stability against small perturbations of the game. Let us begin by considering some ways of eliminating unreasonable actions.

Given a strategic-form game, we say that an action c_i for player i is <u>strongly dominated</u> if there exists some some randomized strategy σ_i such that player i would always get a strictly higher expected payoff from σ_i than from c_i, no matter what actions are used by the other players; that is, for every $(c_1,\ldots,c_{i-1},c_{i+1},\ldots,c_n)$,

$$u_i(c_1,\ldots,c_i,\ldots,c_n) < \sum_{d_i} \sigma_i(d_i)\, u_i(c_1,\ldots,d_i,\ldots,c_n).$$

An action c_i is <u>weakly dominated</u> if there exists some randomized strategy σ_i such that player i would never get a strictly lower expected payoff from σ_i than from c_i, and could possibly get a strictly higher expected payoff from σ_i than from c_i, depending on what actions are used by the other players.

It seems unreasonable to suggest that a player should use an action that is strongly dominated, since he can surely expect better with the dominating strategy. Similarly, equilibria that involve weakly dominated actions may be considered unreasonable. For example, consider Example 4 above and Example 5 below:

		Player 2	
		L	R
Player 1	T	5,5	0,5
	B	5,0	1,1

Example 5

In each of these examples, there are two equilibria ((T,L) and (B,R)), but the action L is weakly dominated by R. So, by the criterion of elimination of dominated actions, (B,R) is the only reasonable equilibrium in both examples. Notice that, in Example 5, the "more reasonable" equilibrium is actually worse for both players.

After the (weakly or strongly) dominated actions have been eliminated from a game, the smaller game that remains may have new dominated actions. Luce and Raiffa [1957] suggested we should continue to eliminate dominated actions iteratively until there are no dominated actions in the game that remains. For example, iterative elimination of weakly dominated actions leaves only the equilibrium at (z_1,z_2) in the following game:

		Player 2		
		x_2	y_2	z_2
	x_1	3,3	0,3	0,0
Player 1	y_1	3,0	2,2	0,2
	z_1	0,0	2,0	1,1

Example 6

(The order of elimination is x_1 and x_2 first, then y_1 and y_2.)

Several related criteria for identifying unreasonable actions have been suggested. Harsanyi [1975] proposed a concept of inferior actions, Bernheim [1984] and Pearce [1984] proposed a concept of unrationalizable actions, and Myerson [1984c] proposed a concept of codominated actions. Each of these concepts includes all of the weakly dominated actions. One might argue that an equilibrium should be considered "unreasonable" if it uses actions that can be eliminated (or iteratively eliminated) by any of these concepts.

Concepts of sequential rationality are applied to dynamic games, in extensive or multistage form. As discussed in section 2, any dynamic game can be reduced to an equivalent strategic-form game. The Nash equilibria of a game in extensive or multistage form are defined to be the Nash equilibria of the equivalent strategic-form game. Unfortunately this definition admits too many equilibria, including some that are clearly irrational.

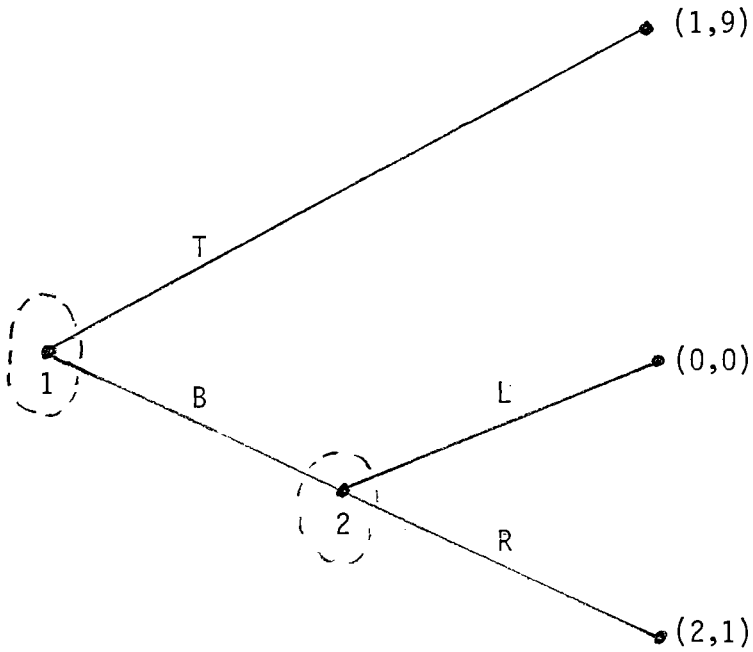
When a decision maker chooses his plan of action in advance, the maximization of expected utility (as viewed ex ante) does not impose any restrictions on what he should plan to do after observing an event that has probability zero. Thus, when we compute the Nash equilibria from the strategic-form reduction of a dynamic game, we abandon all rational restrictions on players' behavior in events that have probability zero. This might not sound like a serious problem, since a zero-probability event should (almost) never occur. However, in game theory (unlike probability theory), the zero-probability events are determined endogenously by the plans or strategies of the players, so we cannot simply ignore such events a priori. An event that has probability zero in one equilibrium may have positive probability in another. So we want to identify the equilibria in which every player is behaving rationally in all events, not just in the positive-

probability events. To do so, we must analyze dynamic games directly in extensive or multistage form, not just in the strategic-form reduction.

For such analysis, Kreps and Wilson [1982] proposed a concept of sequential equilibrium, which refined Selten's [1975] earlier definition of subgame-perfect equilibrium. To characterize a sequential equilibrium, we must specify, not only the action or randomized strategy that each player would use at each stage in each of his possible information states, but also the beliefs that he would have at each stage in each of his information states (including states with probability zero). In every information state, the designated strategies should be rational, in the sense that they maximize the player's conditionally expected payoff given his beliefs (about the other players and chance events) at this information state. The designated beliefs at the various information states should be consistent with each other and with the designated strategies, according to the rule of Bayesian inference from probability theory. A sequential equilibrium is any such rational and consistent designation of strategies and beliefs for all players in all information states. (See Kreps and Wilson [1982] for a more precise definition.)

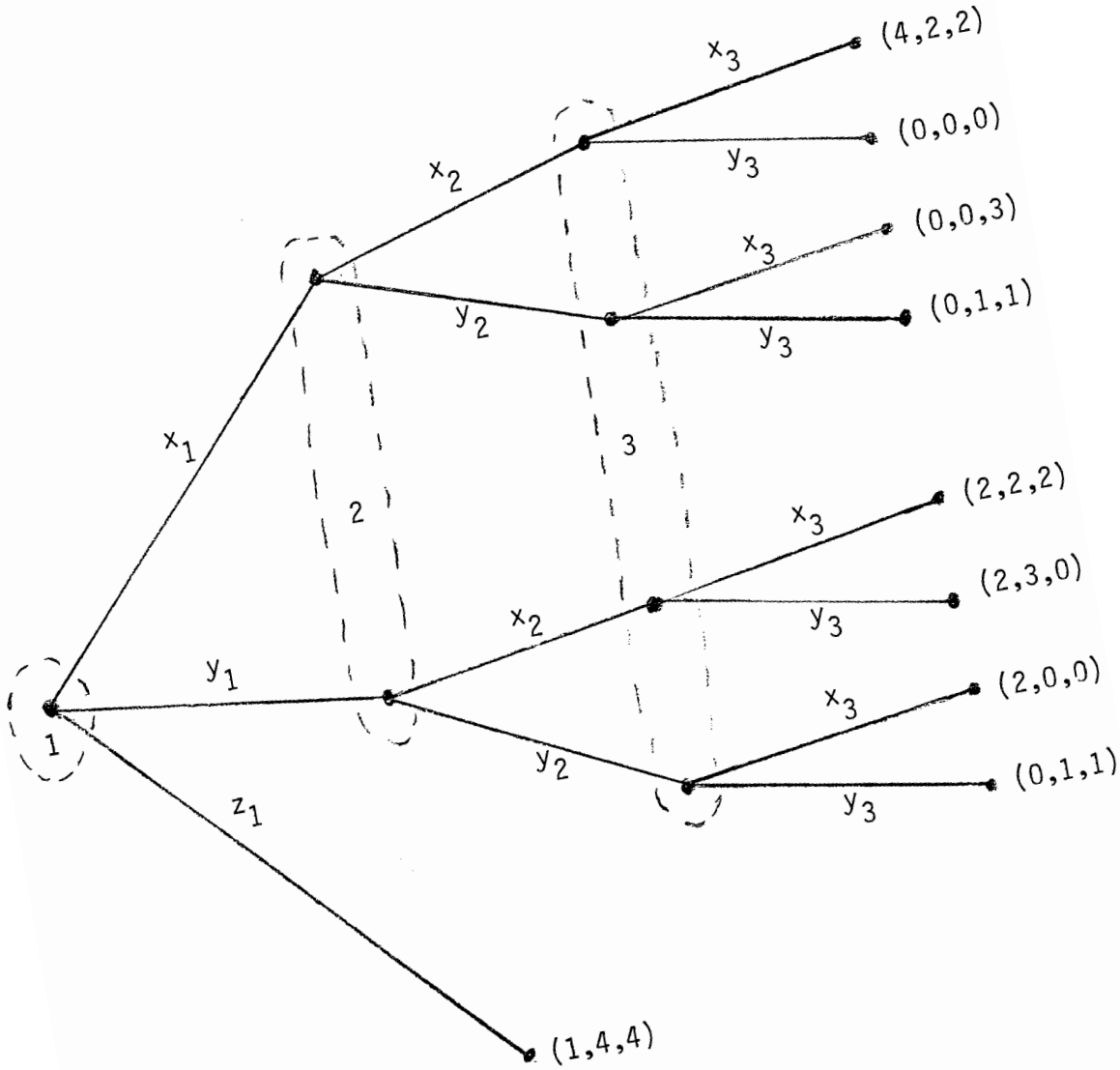
[INSERT FIGURE 1 (EXAMPLE 7) ABOUT HERE]

Consider Example 7, which is shown in extensive form in Figure 1. In this game, player 1 first chooses between T and B. If player 1 choses T, then the payoffs (u_1,u_2) are (1,9), independent of any actions by player 2. If player 1 chooses B then player 2 is informed of this fact and must choose between L and R. The payoffs (u_1,u_2) are (0,0) after B and L, and are (2,1) after B and R. The equivalent strategic-form game is just Example 4, which



Example 7.

FIGURE 1



Example 8.

FIGURE 2

has two Nash equilibria: (T,L) and (B,R). Let us consider the (T,L) equilibrium. Player 1 would prefer T if he expected player 2 to choose L after B; and, at the beginning of the game, player 2 would be willing to plan to choose L after B if he were sure that player 1 would choose T (since the plan would never have to be used). Thus, (T,L) is a Nash equilibrium. But if player 2 cannot actually precommit himself to the L-if-B plan at the beginning of the game, then we must ask what player 2 would rationally choose if he were in the position of choosing between L and R after observing B. In such a circumstance, player 2 should certainly choose R (giving him a payoff of 1) rather than L (giving 0). Thus, (T,L) is not a sequential equilibrium of Example 7. The unique sequential equilibrium is (B,R).

Of course, elimination of dominated actions already excluded (T,L) from the set of reasonable equilibria of Example 4. However, more complicated games, such as Example 8 below, may have Nash equilibria that are not sequential equilibria, even though there are no dominated actions.

[INSERT FIGURE 2 (EXAMPLE 8) ABOUT HERE]

In Example 8, player 1 first chooses among x_1, y_1, and z_1. If he chooses x_1 or y_1, then player 2 must choose between x_2 and y_2, and player 3 must choose between x_3 and y_3. The dotted curves indicate that, when players 2 and 3 make these choices, they do not observe each others' choices, and they do not observe whether x_1 or y_1 was chosen by player 1, but they do know that player 1 has not chosen z_1. If player 1 chooses z_1, then the final payoffs (u_1,u_2,u_3) will be (1,4,4), without players 2 and 3 making any choices at all. Otherwise, the three players' payoffs (u_1,u_2,u_3) depend on the actions of all three players, as indicated at the right ends of the tree.

There are two nonrandom Nash equilibria of this game, (x_1,x_2,x_3) and (z_1,y_2,y_3); but only (x_1,x_2,x_3) is a sequential equilibrium. To try to justify (z_1,y_2,y_3) as a sequential equilibrium, we would have to specify what players 2 and 3 should believe about player 1's choice (x_1 or y_1) if they observed that he did not choose z_1 as expected. If player 2 would believe x_1 and player 3 would believe y_1 under such circumstances then their actions would be rational, but such beliefs would obviously not be consistent. It can be shown that, for any probability distribution over x_1 and y_1, either player 2 or player 3 could expect to gain by deviating from (y_2,y_3). Thus, (z_1,y_2,y_3) is not a sequential equilibrium.

On the other hand, suppose that we revise Example 8 by setting $u_2(y_1,x_2,y_3)$ and $u_3(x_1,y_2,x_3)$ both equal to some number α, where $\alpha \leqslant 2$. (In Figure 2, we had $\alpha = 3$.) Then (z_1,y_2,y_3) can be supported as a sequential equilibrium, by specifying that players 2 and 3 would both assign probability 1/2 to x_1 and probability 1/2 to y_1 if they unexpectedly learned that z_1 was not chosen.

The concept of sequential equilibrium gives us a stronger characterization of rational behavior in dynamic games than Nash equilibrium does. However, we pay an analytical price for changing our solution concept to sequential equilibrium, because we can no longer restrict our attention to the strategic form. Two dynamic games that both reduce to the same equivalent game in strategic form may have different sets of sequential equilibria.

Nevertheless, there is still strong interest in studying refinements of the Nash equilibrium concept for games in strategic form. One desirable property that we might want such a refinement to satisfy is that any equilibrium that it accepts should correspond to a sequential equilibrium in every dynamic game that can be reduced to the given strategic-form game. This

property has been proven by Selten [1975] for his concept of perfect equilibrium, by Kohlberg and Mertens [1983] for their concept of stable equilibria, and by Van Damme [1984] for Myerson's [1978a] concept of proper equilibrium. (Selten [1975] does not use the normal reduction to strategic form, however. He represents each of a player's information states in each stage by a different "agent" in his strategic-form reduction.)

The definitions (omitted here) of perfect, proper, and stable equilibria are all motivated by the general idea that a reasonable equilibrium ought to be stable (in some sense) when the game is slightly altered by introducing small probabilities of players' mistakes or small perturbations of the payoffs. Kalai and Samet [1982] defined a concept of persistent equilibrium, which is also derived from a concern for such stability.

Harsanyi and Selten [1985] have considered ways to try to identify one equilibrium that would be most rational, in some sense, for every finite game in strategic form. However, they have also shown that it is impossible to select a unique equilibrium for every game in a way that depends continuously on the payoffs. This results casts doubt on whether we could ever hope for a truly satisfactory general solution to the problem of multiple equilibria.

5. Extensions of the Equilibrium Set

There are many games, like the Prisoner's Dilemma (Example 2), in which the Nash equilibria yield very low payoffs for the players, relative to other non-equilibrium outcomes. In such situations, the players would want to transform the game, if possible, so as to extend the set of equilibria to include better outcomes. We consider here three such ways that a game might be transformed: with contracts, repetition, and communication.

Joint contracts are the simplest way to extend the equilibrium set. For example, in the Prisoner's Dilemma (Example 2), the players might consider signing a contract that says: "We, the undersigned, promise to choose actions T and L, unless this contract is signed by only one player, in which case he will choose B or R." The option to sign this contract may be introduced into the game description (as action "S"); and then the transformed game

		Player 2		
		L	R	S
	T	5,5	0,6	0,6
Player 1	B	6,0	1,1	1,1
	S	6,0	1,1	5,5

has an equilibrium at (S,S), which gives a payoff of 5 to each player.

In general, given any strategic-form game, a <u>correlated strategy</u> for a set of players is any probability distribution over the set of all possible combinations of actions that they might choose. The <u>minimax value</u> (or <u>security level</u>) for a player i is the best expected payoff that he could get against the worst (for him) correlated strategy that the other (n - 1) players could use. That is, the minimax value for player 1 would be

$$\underset{\sigma}{\text{minimum}} \left(\underset{c_1}{\text{maximum}} \sum_{(c_2,\ldots,c_n)} \sigma(c_2,\ldots,c_n)\, u_1(c_1,c_2,\ldots,c_n)\right)$$

where σ ranges over the set of all probability distributions on $C_2 \times \ldots \times C_n$. The correlated strategy that achieves this minimum is called

the minimax strategy against player 1.

Let μ be a correlated strategy for all the players. The expected payoff to i from μ is

$$U_i(\mu) = \sum_{(c_1,\ldots,c_n)} \mu(c_1,\ldots,c_n)\, u_i(c_1,\ldots,c_n),$$

where $\mu(c_1,\ldots,c_n)$ denotes the probability assigned to the combination of actions $(c_1,\ldots,c_n)$ by the correlated strategy μ. To implement μ, the players might use a trustworthy mediator (or a computer with a random number generator) to randomly designate a combination of actions in $C_1 \times \ldots \times C_n$ according to these probabilities. Our basic question is, under what circumstances might the players all voluntarily sign a contract that commits them to implement μ? Obviously, no player would sign if his expected payoff from μ were less than his minimax value (since he can always get at least his minimax value, whatever the other players might do). Conversely, suppose that each player's expected payoff from μ is greater than or equal to his minimax value; then it would be a Nash equilibrium for every player to sign a contract that says: "We the undersigned agree to choose the actions designated for us randomly according to the correlated strategy μ, unless this contract is signed by all but one player, in which case we will implement the minimax strategy against that player." (We assume in this argument that each player signs independently, without knowing which other players have signed.) Thus, if players' actions can be regulated by joint contracts, any correlated strategy that gives every player at least his minimax value can be implemented by a Nash equilibrium of the transformed game with contract-signing.

Of course, there are many situations in which the players cannot commit themselves to binding contracts. The players' actions might be unobservable

to the legal enforcers of contracts; or sanctions to guarantee compliance with contracts might be inadequate; or some players' actions might be inalienable rights (such as the right to quit a job).

The effect of repeating a game is very similar to the effect of allowing binding contracts. Any correlated strategy that gives each player a higher expected payoff than his minimax value can be enforced in an equilibrium of an infinitely repeated game, if each player's objective is to maximize his long-run average payoff per round. The essential idea is that, if any one player diverged from his role in the correlated strategy then the others would punish him with the minimax strategy against him for many rounds of the game, and any player who diverged from his designated role in punishing another player would similarly be liable for such punishment, so the punishment process is self enforcing. This idea is known as the "Folk Theorem" because it was discussed informally for several years before a rigorous formulation and proof was given by Rubinstein [1979]. For a general introduction to the study of repeated games, see Aumann [1981].

The effect of allowing players to communicate in a strategic-form game, without binding contracts and without repetition, was first studied by Aumann [1974], who defined the concept of <u>correlated equilibrium</u> for such games. To understand this concept, let us begin by considering Example 9.

		Player 2	
		L	R
Player 1	T	5,1	0,0
	B	4,4	1,5

Example 9

There are three equilibria of this game: (T,L), giving payoffs $(u_1,u_2) = (5,1)$; (B,R), giving payoffs (1,5); and a randomized equilibrium giving expected payoffs (2.5,2.5). The best symmetric payoffs (4,4) cannot be achieved by the players without binding contracts, because (B,L) is not an equilibrium. (Player 1 would choose T if he expected player 2 to choose L.) The expected payoffs (3,3) can be achieved by the players, with communication but without binding contracts, by tossing a coin and planning to choose (T,L) if heads and (B,R) if tails. Such a plan is self-enforcing, even though the coin has no binding force on the players, because neither player could gain by unilaterally diverging from the plan.

Even better, $(3\frac{1}{3},3\frac{1}{3})$ can be achieved in this game, with the help of a mediator. Suppose that a mediator randomly recommends actions to the two players in such a way that each of the pairs (T,L), (B,L), and (B,R) may be recommended with probability 1/3. Suppose also that each player hears only his own recommended action from the mediator. Then, even though the mediator's recommendation has no binding force, it is a Nash equilibrium (of the transformed game with such mediated communication) for both players to plan to obey the mediator's recommendations. If player 1 hears a recommendation "B," then he thinks that player 2 may have been told to do L or R with equal probability, in which case his expected payoff from B (2.5) is as good as from T. If player 1 hears "T" then he knows that player 2 was told to do L, in which case T is player 1's best action. So player 1 is willing to obey the mediator if he expects player 2 to obey, and a similar argument applies to player 2. Randomizing equally between (T,L), (B,L) and (B,R) with equal probability gives expected payoffs $(3\frac{1}{3},3\frac{1}{3})$.

More generally, for any strategic-form game (as in Section 2), Aumann

[1974] defined a correlated equilibrium to be any correlated strategy that can be achieved as an equilibrium with the help of such a mediator. Formally, a correlated equilibrium is any probability distribution μ over $C_1 \times \ldots \times C_n$ such that, for every player i and every function δ_i that maps C_i into C_i,

$$U_i(\mu) \geq \sum_{(c_1,\ldots,c_n)} \mu(c_1,\ldots,c_n)\, u_i(c_1,\ldots,\delta_i(c_i),\ldots,c_n).$$

That is, no player i should expect to gain from disobeying the mediator's recommendations by any rule δ_i (doing $\delta_i(c_i)$ when told to do c_i), if the mediator's recommendations are randomly selected according to the probability distribution μ and every other player is expected to obey his recommendations. (Player i's prospective disobedience $\delta_i(c_i)$ can depend only on c_i, the recommended action for player i, because i does not hear the mediator's recommendations to the other players.) These inequalities are called strategic incentive constraints (or moral-hazard constraints), because they constrain the set of correlated strategies that a mediator could implement without giving any player an incentive to disobey. By analyzing these incentive constraints, we can characterize the set of all correlated equilibria for a game. For example, it is straightforward to show that $(3\frac{1}{3}, 3\frac{1}{3})$ is the best symmetric payoff allocation that can be achieved by any correlated equilibrium for Example 9. (See Myerson [1985a].)

For Bayesian games with incomplete information, communication would allow the players' actions to depend on each others' types, as well as on extraneous random variables like coin tosses. Formally, μ is a (direct) communication mechanism (or a generalized correlated strategy) for the players in a Bayesian game if, for every possible combination of types $(t_1,\ldots,t_n)$, μ specifies a probability distribution, denoted $\mu(\cdot|t_1,\ldots,t_n)$, over the set of possible

combinations of actions $(C_1 \times \ldots \times C_n)$. We let $\mu(c_1,\ldots,c_n|t_1,\ldots,t_n)$ denote the probability that each player i should be told to do c_i, if every player j reported his type to be t_j, in the communication mechanism μ. We let $U_i(\mu|t_i)$ denote the expected payoff to player i from the communication mechanism μ when his type is t_i; that is

$$U_i(\mu|t_i) = \sum_{t_{-i}} p_i(t_{-i}|t_i) \sum_c \mu(c|t)\ u_i(c,t).$$

(We use here the notation $t_{-i} = (t_1,\ldots,t_{i-1},t_{i+1},\ldots,t_n)$, $c = (c_1,\ldots,c_n)$, and $t = (t_1,\ldots,t_n)$.)

Suppose that a mediator were to help the players implement a communication mechanism μ in a Bayesian game. Suppose also that the mediator can communicate confidentially with each player, receiving type reports and sending action recommendations; but he cannot force the players to report honestly or to act obediently. Then honest reporting and obedient action by all players would be a Bayesian equilibrium if and only if μ satisfies the following general <u>incentive constraints</u>, for every player i, every pair of possible types t_i and s_i, and every function δ_i from C_i into C_i:

$$U_i(\mu|t_i) \geq \sum_{t_{-i}} p_i(t_{-i}|t_i) \sum_c \mu(c|t_{-i},s_i)\ u_i((c_{-i},\delta_i(c_i)),t).$$

(Here we use the notation $(t_{-i},s_i) = (t_1,\ldots,t_{i-1},s_i,t_{i+1},\ldots,t_n)$, and $(c_{-i},\delta_i(c_i)) = (c_1,\ldots,c_{i-1},\delta_i(c_i),c_{i+1},\ldots,c_n)$.) That is, no player i should expect to gain, when his type is t_i, by reporting type s_i and then disobeying according to δ_i, when the mediator chooses his recommendations (as a random function of the players' reports) according to μ and when all other players are expected to be honest and obedient. Any communication mechanism

that satisfies these incentive constraints may be called a communciation equilibrium (or a generalized correlated equilibrium, or an incentive-compatible mechanism) for the players in the Bayesian game.

The above definition of communication equilibria was motivated by considering what can be achieved by a mediator with a fully centralized communication system. However, the class of communication equilibria defined above actually characterizes what can be achieved by any kind of communication system, in the following sense. Any communication system effectively transforms any given Bayesian game into a new Bayesian game, in which each player's "action" is a communication strategy (that is, a specification of what messages he will send and how he will choose his action in the originally given game as a function of the messages that he receives). For any communication system and any Bayesian equilibrium of the transformed game with communication, we can construct an equivalent communication equilibrium μ (as above) by letting $\mu(c_1,\ldots c_n|t_i,\ldots,t_n)$ to be the probability that $(c_1,\ldots,c_n)$ would be the actions chosen in the original game if $(t_i,\ldots t_n)$ were the players' types, when the players use their equilibrium communication strategies. It is straightforward to show (see Myerson [1982]) that this μ must satisfy the incentive constraints from the preceding paragraph. This result is called the revelation principle, since it shows that there is no loss of generality in only considering communication systems in which all players reveal all their information to a central mediator.

By the revelation principle, for any social welfare function, the problem of designing an optimal communication system can be solved as a mathematical optimization problem, where the incentive constraints define the feasible set. These constraints are linear in μ, so this optimization problem can often be explicitly solved by well-known techniques. In fact, it is sometimes

easier to characterize the set of communication equilibria of a game than the set of Bayesian equilibria of the same game without communication, even though the communication equilibria include the Bayesian equilibria. The set of Bayesian equilibria has no simple geometrical structure, but the set of communication equilibria is a convex polyhedron.

At this point, it may be helpful to distinguish between the terms mediator and arbitrator as they have been used in this paper. Both terms refer to an outside individual who intervenes in a game to help the players in some way. A mediator acts as a communication channel between the players, thereby transforming the game and enlarging the set of equilibria. An arbitrator (as described at the end of Section 3) helps to determine which equilibrium should be implemented by the players in a game with multiple equilibria. In short, a mediator is an outside individual who communicates with the players in order to enlarge the set of equilibria, whereas an arbitrator is an outside individual who uses his authority or power of suggestion to help the players to select among multiple equilibria. Of course, there are many situations in which an individual may serve both as a mediator and as an arbitrator in a game, but these two functions are logically distinct.

To appreciate this distinction, suppose that the players in some game are in communication with many different mediators, each of whom uses a different incentive-compatible mechanism for determining his recommendations. There are many equilibria of this transformed game with communication. For example, the players could simply babble to all the mediators and implement some Bayesian (or Nash) equilibrium of the original game without communication. (We may say that a player babbles to a mediator if the player's report to the mediator is chosen independently of the player's type and the player's action is chosen

independently of the mediator's recommendation. We may suppose that a babbling player simply randomizes uniformly over the set of possible type-reports.) It is also an equilibrium for the players to be honest and obedient to one particular mediator and babble to all the others. An arbitrator could designate any one of the mediators as the one whom the players should obey.

In some situations, a mediator may also have the power to act as an auditor or as a regulator of the players in a game. A regulator is an individual who can directly control the players' actions, so that they cannot disobey him. (When we considered the effect of binding contracts, we implicitly assumed the existence of a regulator.) If a mediator is also a regulator, then he can implement any communication mechanism that satisfies the following informational incentive constraints (or self-selection constraints) for every player i and every pair of possible types t_i and s_i:

$$U_i(\mu|t_i) \geq \sum_{t_{-i}} p_i(t_{-i}|t_i) \sum_c \mu(c|t_{-i},s_i)\, u_i(c,t).$$

(That is, the mediator only needs to guarantee that no player could expect to gain by lying about his type.) An auditor is an individual who can directly observe and verify the players' types, so that they cannot lie to him. A mediator who can both regulate and audit the players can implement any communication mechanism, without regard for incentive constraints.

Dynamic multistage games with communication have been studied by Myerson [1984d]. The set of communication equilibria for such games are defined by incentive constraints that are similar to those for strategic-form and Bayesian games. Unfortunately, the set of communication equilibria of a multistage game cannot be identified with the set of correlated equilibria of the "equivalent" strategic-form game (as in Section 2), because opportunities

for the players to communicate after the beginning of the game would be suppressed in the reduction to strategic form. To characterize sequential rationality in multistage games with communication, a concept of sequentially rational communication equilibria has also been defined. Myerson [1984d] showed that a communication equilibrium of a multistage game is sequentially rational if and only if it never involves the use of certain sequentially codominated actions, which include dominated actions. Thus, two approaches to the refinement of the Nash equilibrium concept (sequential rationality and elimination of unreasonable actions) can be unified in the context of multistage games with communication.

6. The Nash Bargaining Solution

We have seen that the problem of multiple equilibria can create the role of an arbitrator, an outside individual who can, by his authority or power of suggestion, determine which equilibrium will be implemented. If there is no arbitrator and no other external determinant of a focal equilibrium, then one prestigious player (called the principal) might have a similar authority to select among the equilibria. But another possibility is that the players may jointly determine the equilibrium to be implemented, by some process of pre-play bargaining or negotiation. That is, the focal equilibrium could be determined by a consensus among all the players, where the consensus is reached through negotiations in which every player has an opportunity to participate. The fundamental problem of cooperative game theory is to predict the negotiated focal equilibria that might be selected by such a process. The second half of this paper (Sections 6-9) is an introduction to the ideas of cooperative game theory.

Let us begin by considering Example 10, the "Divide the Dollars" game. In this game, there are two players who can divide $100 between themselves, provided that they can both agree on the division; otherwise they each get nothing. To be specific, let us suppose that each player simultaneously chooses a demand, which is any number between 0 and 100. If the sum of their demands is less than or equal to 100, then each player gets a payoff equal to his demand. If the sum of the demands is greater than 100, then each player gets a payoff of zero. (This example is a special case of the demand games studied by Nash [1953]. In this case we are identifying utility with money.)

This game has multiple equilibria. For any number x such that $0 < x < 100$, there is an equilibrium in which player 1 demands x and player 2 demands $100 - x$; so every efficient allocation is achievable in an equilibrium. There are also inefficient randomized equilibria. For example, there is an equilibrium in which each player (independently) randomizes between demanding 1, with probability 1/99, and demanding 99, with probability 98/99, so that the probability that both get nothing is almost 98%.

In Section 3, we argued that the outcome of such a game with multiple equilibria is likely to be a focal equilibrium which may be designated by an arbitrator or other environmental factors. Since "environmental factors" are not included in the given mathematical description of the game, analysis of such focal equilibria may be beyond the scope of mathematical game theory. For example, if player 1 is male and player 2 is female, and if there has been a tradition that males take 75% of mutually-feasible gains in the players' society, then the (75,25) equilibrium is the focal and most likely outcome of this game. On the other hand, if the selection of a focal point depends on an arbitrator, rather than on some exogenous social tradition, and if we assume that an arbitrator's judgement should depend only on the given mathematical

description of the game, then we can hope to apply the methods of mathematical game theory to predict the outcome of the game. To do so, we must develop a theory of arbitration in games.

For the Divide the Dollars game, it seems clear that an impartial arbitrator should recommend the equilibrium in which each player gets 50. We should expect that an impartial arbitrator would recommend a symmetric allocation, because players 1 and 2 are completely symmetric in the mathematical description of the game. The (50,50) equilibrium is the only symmetric equilibrium in which the players are sure to divide all the available money. Thus, the (50,50) equilibrium is likely to be recommended by an arbitrator because it is the unique outcome that is both equitable and efficient.

Once (50,50) has been identified as the impartially arbitrated solution, it is not really necessary to have an arbitrator actually present when the game is played. The two players, being intelligent, can predict arbitrated settlements as well as we theorists can. Thus, it should be common knowledge among the players that an impartial arbitrator would recommend the (50,50) equilibrium, and this fact gives the (50,50) equilibrium an intrinsic focal property, even when no arbitrator is present. That is, <u>properties of equity and efficiency can determine a focal equilibrium in a game, as well as social or environmental factors</u>. Thus, in the Divide the Dollars game, unless there is some strong social tradition pointing to some other outcome, the (50,50) equilibrium is the most likely outcome to be chosen by the players, even when there is no actual arbitrator.

To extend this analysis to other games, we need a general theory of fair arbitrated settlements in games. The first and most compelling of such theories in the literature (see Roth [1979]) is the bargaining solution of

Nash [1950, 1953].

The Nash bargaining solution can be defined as a function of a feasible set (F) and a disagreement point (v), which in turn depend on the strategic form of the game. The feasible set of an n-player game is a closed and convex subset of $\mathbb{R}^n$, representing the set of all allocations of expected utility that the players can jointly achieve. If the players can make jointly binding contracts to regulate their actions, then we may define the feasible set F to be the set of all vectors $(U_1(\mu),\ldots,U_n(\mu))$ such that μ is any correlated strategy. If the players cannot make jointly binding contracts, then we may define the feasible set F to be the set of all vectors $(U_1(\mu),\ldots,U_n(\mu))$ such that μ is any correlated equilibrium, satisfying the strategic incentive constraints.

The disagreement point of an n-player game is a vector in $\mathbb{R}^n$ that represents the utility payoff that each player could guarantee himself if he did not coordinate with the other players. One way to formalize this idea is to define the disagreement point to be the vector $v = (v_1,\ldots,v_n)$ where each v_i is the minimax value for player i. An alternative suggestion is to let v be the vector of expected utility payoffs that the players would get in some focal Nash equilibrium. A third suggestion, developed by Nash [1953] is to define v to be the vector of expected utility payoffs that the players would get if each carried out some (endogenously determined) optimal threat. (The distinction between these definitions corresponds to different assumptions about whether the players can commit themselves before arbitration to offensive and defensive threats that would be implemented if some player subsequently refused to accept the arbitrated settlement.)

Once a feasible set and a disagreement point have been specified, it seems reasonable that one should be able to define an equitable and efficient

payoff allocation as a function of these two structures, without further reference to the underlying strategic-form game. The efficient payoff allocations, in the sense of Pareto, are precisely the points on the upper boundary of the feasible set. That is $(x_1,\ldots,x_n)$ is <u>efficient</u> if it is in the feasible set and there is no other vector $(y_1,\ldots,y_n)$ in the feasible set such that $y_i \geqslant x_i$ for all i, with at least one strict inequality. The definition of equitable allocations is more problematic, but equity is supposed to mean that each player's gains from the arbitrated settlement are in some sense commensurate with every other player's gains. Once the disagreement point is specified, to represent the consequences of rejecting arbitration, the utility gains from arbitration can be computed for each player at each feasible payoff allocation.

Let us assume henceforth that there is a point in the feasible set in which every player does strictly better than in the disagreement point, so that no player would want to force disagreement.

Nash [1950] listed several axioms that an impartial arbitration procedure should satisfy. If the players are symmetric, as in the Divide the Dollars game, then they should get equal payoffs. The payoff allocation selected by the procedure should be on the efficient boundary of the feasible set, and should give each player a higher payoff than he gets at the disagreement point. If a game is changed in such a way that the feasible set is made smaller, but the disagreement point is unchanged and the old arbitrated settlement is still feasible, then the new arbitrated settlement should be the same as the old (because the lost utility allocations would not have been used anyway, so they are irrelevant). If a game is changed by multiplying one player's utility function by a positive constant or by adding a constant, then his payoff in the settlement should be changed by the same multiplicative or

additive constant, and all other players' payoffs should remain the same. (This is because multiplying by a positive constant or adding a constant does not change any of the decision-theoretic properties of a utility function. Thus, a player's utility function can only be defined up to such linear transformations.) Nash's remarkable result is that these properties are satisfied by only one arbitration rule: choose the utility allocation $(x_1,\ldots,x_n)$ that maximizes

$$\prod_{i=1}^{n} (x_i - v_i),$$

the multiplicative product of the players' gains over the disagreement point $(v_1,\ldots,v_n)$, subject to the constraints that $(x_1,\ldots,x_n)$ is in the feasible set and $x_i \geq v_i$ for every i. This allocation is the <u>Nash bargaining solution</u> of the game.

An alternative characterization of the Nash bargaining solution may clarify in what sense it is equitable. We use here the fact that no decision-theoretic properties are affected by multiplying a player's utility function by a positive constant. Thus, any weighted-utility function $\lambda_i u_i$, where $\lambda_i > 0$, can represent player i's preferences as well as the given utility function u_i. This fact creates a problem for the arbitrator who wants to treat the players equitably in his recommended settlement. Since there is no decision-theoretic basis for distinguishing between the different weighted-utility functions as representations of a given player's preferences, which functions should be used in making the interpersonal comparisons that equity requires?

There are actually two kinds of interpersonal comparisons of utility that people often try to make in games: utilitarian comparisons and egalitarian

comparisons. Utilitarian comparisons are implicit in the sentence: "You should do this for me because it will help me more than it hurts you." A utilitarian optimum is a feasible outcome that maximizes the sum of all players' utilities. Egalitarian comparisons are made implicit in the sentence: "You should do this for me because I am doing more for you." An egalitarian optimum is an outcome in which all players gain equally over the disagreement point. In general, there may be nothing that is optimal in both senses at once. Furthermore, if we make comparisons in <u>weighted</u> utility scales, the sets of optimal outcomes change as the λ_i weights are changed. However, there always exists some vector of weights $(\lambda_1,\ldots,\lambda_n)$ such that, when we make interpersonal comparisons in terms of weighted utilities, the utilitarian optima and the egalitarian optima intersect; and this intersection is exactly the Nash bargaining solution. That is, given the feasible set F and the disagreement point $(v_1,\ldots,v_n)$, an allocation $(x_1,\ldots,x_n)$ in F is the Nash bargaining solution if and only if there exists some vector of positive weights $(\lambda_1,\ldots,\lambda_n)$ such that

$$\sum_{i=1}^{n} \lambda_i x_i = \underset{y \text{ in } F}{\text{maximum}} \sum_{i=1}^{n} \lambda_i y_i$$

(where $y = (y_1,\ldots,y_n)$) and

$$\lambda_1 x_1 - \lambda_1 v_1 = \lambda_2 x_2 - \lambda_2 v_2 = \ldots = \lambda_n x_n - \lambda_n v_n.$$

We may refer to the λ_i that satisfy these conditions as the <u>natural utility weights</u> for the given game. Thus, the Nash bargaining solution gives the players equal gains over the disagreement point, in terms of the naturally weighted utility scales for the game.

Nash [1950] stated that his bargaining solution was intended to predict the outcome of bargaining, without an arbitrator, between two players who have

equal ability to bargain and who can jointly select any feasible correlated strategy for the given game. ("Feasible" here may mean either subject to incentive constraints or not, depending on whether binding contracts are possible.) As in the Divide the Dollars game, we can make the logical transition from the theory of arbitration to the theory of unarbitrated bargaining by invoking the focal-point effect. Thus, for any two-person bargaining problem, we may predict that the equity and efficiency properties of the Nash bargaining solution will lead the players to select it in their bargaining process, unless some other environmental factor or tradition focuses more strongly on some other outcome.

There may be situations in which players have unequal bargaining ability. To describe such situations, nonsymmetric versions of the Nash bargaining solution have been proposed. Nonsymmetric bargaining solutions may also be applied in arbitration, when the arbitrator feels that one player's welfare deserves relatively more weight, because of the player's intrinsic personal characteristics. For example, if player 1 is single and player 2 represents a family of four people, then an arbitrator in the Divide the Dollars game might recommend the 20-80 division, to equalize per capita gains.

In general, a <u>nonsymmetric Nash bargaining solution</u> may be defined as any solution function (mapping feasible sets with disagreement points into payoff allocations) that satisfies all of Nash's axioms except the axiom of symmetry. It can be shown (see Kalai [1977]) that a nonsymmetric Nash bargaining solution always maximizes some product of the players' gains raised to various powers; that is, it maximizes

$$\prod_{i=1}^{n} (x_i - v_i)^{\alpha_i}$$

subject to $(x_1,\ldots,x_n)$ being in the feasible set F and satisfying $x_i \geqslant v_i$ for all i. Here the exponents $(\alpha_1,\ldots,\alpha_n)$ are some nonnegative parameters (not all zero) that represent the players' relative weight in arbitration or relative bargaining ability.

The above discussion has followed the cooperative approach to the theory of bargaining. There is an alternative noncooperative approach to the theory of bargaining, which was originally advocated by Nash [1951] himself in his seminal paper on equilibria. (So it is also called Nash's program.)

The noncooperate approach to bargaining is to try to explicitly describe the sequence of decisions and actions that individual players can make during the bargaining process. Each player's role in the process of "jointly selecting a correlated strategy" must be made through some sequence of actions that he controls individually (actions of making threats and offers, and accepting or rejecting others' offers). Since the outcome of this process is the selection of a feasible correlated strategy to be used in a given game, and since the expected payoffs from this correlated strategy can be computed (as was discussed in Section 5), the bargaining process itself can (in principle) be modelled as a multistage game. Thus, Nash [1951] argued, we should try to predict the outcome of the bargaining process by modelling this game and analyzing its Nash equilibria.

There are difficulties with the noncooperative approach to bargaining. Bargaining between individuals who can communicate fact to face in a sophisticated language such as English is obviously a much more complex process than the simple bargaining models which theorists can study. Any tractable model must make some simplifications which may seem arbitrary or ad hoc. Furthermore, even in a simple bargaining model, the set of equilibria may be very large, especially if players can make bargaining decisions

simultaneously, or if there is incomplete information on the game. From the perspective of Section 5, we could even say that an ideal bargaining process would transform the game so that all of the feasible correlated strategies of the original game would become alternative equilibrium outcomes. (See Crawford [1983] for a comprehensive development of this idea.) In such a bargaining process, there remains a problem of equilibrium selection, where each equilibrium corresponds to an allocation in the feasible set. This selection problem leads us back to cooperative game theory and concepts such as the Nash bargaining solution.

On the other hand, the Nash bargaining solution is limited as a solution concept by the fact that its relevance to equilibrium selection is based on the focal-point effect. Experimental evidence (see Roth and Shoumaker [1983]) suggests that factors from the sociological environment often have much stronger focal effect than the theoretical properties of the Nash bargaining solution. Furthermore, the nonsymmetric Nash bargaining solutions involve a concept of "relative bargaining ability," which begs deeper explanation or analysis. Thus, there has been a growing interest in exploring the noncooperative approach to bargaining.

The noncooperative models of bargaining studied by Rubinstein [1982] give significant insights into what might determine a player's "relative bargaining ability." In these models, there are two players who can alternately make offers to each other. For a specific example, suppose that the two players (beginning with player 1) alternately offer payoff allocations in the feasible set F, either until one player accepts the other's most recent offer, or until the bargaining terminates in disagreement. Each time that player 2 makes an offer instead of accepting player 1's most recent offer, there is an exogenous probability p_1 that the bargaining process will terminate in disagreement.

(Actually we only need to assume that player 2 gets his disagreement payoff in this event.) Similarly, after the first offer, each time that player 1 makes an offer instead of accepting player 2's most recent offer, there is an exogenous probability p_2 that the bargaining will terminate in disagreement (so that player 1 gets his disagreement payoff). Otherwise, when an offer is accepted, both players get the payoffs specified in the offer. Using the methods of Rubinstein [1982], it can be shown that this game has a unique sequential equilibrium, provided that at least one of the given probabilities (p_1 or p_2) is positive. In this unique sequential equilibrium, player 1's first offer is always accepted by player 2. Furthermore, as Binmore [1981] shows, if we let $p_1 = \varepsilon\alpha_1$ and $p_2 = \varepsilon\alpha_2$ and we then take the limit as ε goes to zero, the accepted equilibrium offer converges to the nonsymmetric Nash bargaining solution with parameters α_1 and α_2.

Thus, the parameter α_i, which measures player i's relative bargaining ability in the nonsymmetric Nash solution, can be interpreted as player i's relative ability to make a credible threat to terminate bargaining if his offer is rejected. The rationale for considering very small exogenous probabilities of such termination is that, in these sequential equilibria, both players would lose if termination actually occurred, so that no one would want to enforce the termination ex post. Player i's bargaining ability thus derives from his ability to introduce at least some infinitesimal doubt in the other player's mind as to whether player i might "irrationally" terminate bargaining if his most recent offer were rejected.

7. Cooperative Games with Transferable Utility

The preceding section discussed the Nash bargaining solution in the context of bargaining games with any number of players. In fact, the Nash bargaining solution was originally advocated by John Nash only for the analysis of two-player games, and this limitation seems appropriate. The essential problem is that the Nash bargaining solution, as defined in the preceding section, ignores the possibility that the players might form any effective coalitions among themselves other than the grand coalition that contains all the players together.

To illustrate, let us compare Examples 11 and 12, each a game among three players. In Example 11, the players get payoffs of \$0 unless all three agree on how to divide \$300 among themselves. In Example 12, the rules are the same except that only players 1 and 2 need to agree, to implement the division of the \$300. (Let us equate utility payoffs with dollar payoffs here.) In both games, the disagreement point is (0,0,0), and the feasible set for the grand coalition ($\{1,2,3\}$) is the set of all allocation vectors (x_1,x_2,x_3) such that $x_1 + x_2 + x_3 \leqslant 300$. The difference is that, in Example 11 no coalition of two players could achieve any payoffs other than zero, whereas in Example 12 the coalition $\{1,2\}$ can achieve any allocation vector (x_1,x_2) such that $x_1 + x_2 \leqslant 300$. In Example 11, the Nash bargaining solution (100,100,100) seems reasonable, but it does not seem reasonable in Example 12. When we take account of the fact that players 1 and 2 do not need player 3, the allocation (150,150,0) seems like a much more reasonable outcome for Example 12.

Before we completely dismiss (100,100,100) as an unreasonable prediction for Example 12, let us carefully examine the assumptions implicit in this rejection. To be more explicit, let us suppose that the strategic rules for the second game are that, after nonbinding pre-play communication, players 1

and 2 simultaneously propose an allocation vector (x_1,x_2,x_3) such that $x_1 + x_2 + x_3 \leq 300$; if the proposed vectors are equal then they are implemented; otherwise all three players get zero. For players 1 and 2 to both propose (100,100,100) is an equilibrium of this game, just as (150,150,0) is also an equilibrium. Even if the preplay communication is made an explicit part of the extensive or multistage characterization of the game, there is still an equilibrium in which each of the players 1 and 2 ignores anything that the other player might say (including: "Let's cut out player 3 and both choose (150,150,0)," because each interprets the other's speech as meaningless babble rather than as English), and then both choose (100,100,100). If player 3 has any influence in such matters, he would certaintly want to promote such mutual misunderstanding between players 1 and 2. Thus, the key assumption that we need, to dismiss (100,100,100) as unreasonable, is that players 1 and 2 can <u>negotiate effectively</u> during their pre-play communication opportunities.

In general, when we say that the members of a coalition can <u>negotiate effectively</u>, we mean that, if there were a feasible joint change in the coalition-members' strategies that would benefit them all, then they would actually agree to make such a change unless it contradicted agreements that some members might have made with other (nonmember) players in the context of some other equally effective coalition. The key assumption that distinguishes cooperative game theory from noncooperative game theory may be the assumption that players can negotiate effectively. In our discussion of the Nash bargaining solution, we implicitly assumed that only the grand coalition of all players can negotiate effectively together. In this section we now assume that any coalition or subset of the players can also negotiate effectively.

Because the interactions between $2^n - 1$ different coalitions in an

n-player game can be so complex, a simplifying assumption of transferable utility is often used in cooperative game theory. That is, there is assumed to be a commodity, called money, that players can freely transfer among themselves, such that any player's utility payoff increases one unit for every unit of money that he gets.

With transferable utility, the cooperative possibilities of a game can be described by a characteristic function v that assigns a number v(S) to every coalition S. Here v(S) is called the worth of coalition S and represents the total amount of transferable utility that the members of S could earn together without any help from the other players outside of S. For Example 11, discussed above, the characteristic function is $v(\{1,2,3,\}) = 300$, $v(\{1,2\}) = v(\{1,3\}) = v(\{2,3\}) = 0$, and $v(\{1\}) = v(\{2\}) = v(\{3\}) = 0$. The characteristic function for Example 12 differs from this only in that $v(\{1,2\}) = 300$. (In any characteristic function v, we let $v(\emptyset) = 0$ where $\emptyset$ is the empty set).

Given a game in strategic form, von Neumann and Morgenstern [1944] suggested that the characteristic function should be defined by

$$v(S) = \underset{\sigma_{N-S}}{\text{minimum}} \ \underset{\sigma_S}{\text{maximum}} \left(\sum_{i \text{ in } S} u_i(\sigma_S, \sigma_{N-S}) \right)$$

where σ_S is any correlated strategy for the coalition S, N-S is the set of all players not in S, σ_{N-S} is any correlated strategy for N-S, and $u_i(\sigma_S, \sigma_{N-S})$ is player i's expected utility payoff, before transfers of money, when these correlated strategies are independently implemented. That is v(S) is the maximum sum of utility payoffs that the members of S can guarantee themselves against the best offensive threat for the complementary coalition N-S. It can be shown that such a "minimax" characteristic function is always

superadditive, in the sense that $v(S \cup T) \geqslant v(S) + v(T)$ for any two coalitions S and T such that $S \cap T = \emptyset$. Harsanyi [1963] recommended an alternative way of deriving the characteristic function from the strategic form, such that

$$v(S) - v(N-S) = \min_{\sigma_{N-S}} \max_{\sigma_S} \left(\sum_{i \text{ in } S} u_i(\sigma_S, \sigma_{N-S}) - \sum_{j \text{ in } N-S} u_j(\sigma_S, \sigma_{N-S}) \right).$$

(See also Myerson [1978b] for more on threats and the characteristic function.)

Once the characteristic function of a game has been specified, we can try to predict the outcome of bargaining among the players. Such analysis is based on the assumption, discussed in the preceding section, that the focal bargaining equilibrium will depend on the power structure, rather than on the details of how bargaining proceeds. A player's power is his ability to help or hurt other players by agreeing to cooperate with them or refusing to do so. Thus, the characteristic function is a representation of the power structure in a game.

Let v be any characteristic function of an n-player game with transferable utility. (As usual $N = \{1,2,\ldots,n\}$ denotes the set of all players.) We may say that a coalition S can object to a payoff allocation $x = (x_1,\ldots,x_n)$ if

$$v(S) > \sum_{i \text{ in } S} x_i,$$

so that the members of S could get more together than they get in the allocation x. An allocation x is the core of v if no coalition can object to x and

$$\sum_{i=1}^{n} x_i = v(N),$$

so that x is feasible for the grand coalition N.

The core is a very attractive solution concept, in view of our assumption that any coalition can negotiate effectively. Unfortunately, the core may be empty. Consider Example 13, which differs from Example 11 in that an allocation that is proposed by any two of the three players will be implemented. The characteristic function of this game is $v(\{1\}) = v(\{2\}) = v(\{3\}) = 0$, $v(\{1,2\}) = v(\{1,3\}) = v(\{2,3\}) = 300 = v(\{1,2,3\})$. If player i gets a positive payoff, then the other two players must get less than the \$300, which they can get by themselves, so they can object. Thus, the core of this game is empty. On the other hand, there are some games in which the core is very large. In Example 11, any allocation of the available \$300 is in the core (as long as no player's share is negative).

Shapley [1953] considered the problem of how to select a unique allocation or <u>value</u> for every game represented by a characteristic function. He proposed several natural properties that such a value function should have (linearity, efficiency, symmetric treatment of symmetric players, and zero payoff allocation for powerless "dummy" players), and he showed remarkably that only one value function satisfies these axioms. In this <u>Shapley value</u>, the value assigned to player i in the n-player game represented by v is

$$\phi_i(v) = \sum_{S \subseteq N-i} \frac{|S|!(n - |S| - 1)!}{n!} (v(S \cup \{i\}) - v(S)).$$

(Here $|S|$ is the number of players in S, $n! = 1 \cdot 2 \cdot 3 \cdot \ldots \cdot n$, and N-i is the set of players other than i.) A generalization of the Shapley value to games with infinitely many players has been developed by Aumann and Shapley [1974].

To understand the Shapley value, observe first that it is a linear function from the set of all characteristic functions (which, for n-player games, is a vector space with $2^n - 1$ dimensions) into the set of payoff allocations (n-dimensional vectors). Linearity is composed of two properties:

(1) $\phi_i(\underset{\sim}{0}) = 0,$

(2) $\phi_i(\lambda v + (1 - \lambda)w) = \lambda\phi_i(v) + (1 - \lambda)\phi_i(w),$

for any two characteristic functions v and w and any number λ between 0 and 1. ($\underset{\sim}{0}$ is the characteristic function that assigns worth zero to every coalition; and $\lambda v + (1 - \lambda)w$ is the characteristic function that assigns worth $\lambda v(S) + (1 - \lambda)w(S)$ to each coalition S.) Equation (1) asserts that, if every coalition can only get zero, then each player should get zero. To interpret equation (2), suppose that the players will play tomorrow either a game represented by v, with probability λ, or a game represented by w, with probability $1 - \lambda$. The expected Shapley value to player i is $\lambda\phi_i(v) + (1 - \lambda)\phi_i(w)$, if the players plan to bargain tomorrow. On the other hand, if the players actually bargain today, planning their strategies in advance, then they are playing a game represented by $\lambda v + (1 - \lambda)w$, because today any coalition S can make plans that earn the expected worth $\lambda v(S) + (1 - \lambda)w(S)$. So the value to player i from bargaining today should be $\phi_i(\lambda v + (1 - \lambda)w)$. Equation (2) asserts that it should not matter whether players bargain today (before the resolution of uncertainty) or tomorrow (after the resolution of uncertainty).

Now, let R be any coalition and consider the game w_R defined by

$$w_R(S) = \begin{cases} 1 & \text{if } S \supseteq R, \\ 0 & \text{otherwise.} \end{cases}$$

That is, a coalition that contains all members of R can get a total of one unit of transferable utility; and a coalition that lacks any member of R gets zero. (This game is called the carrier game for R.) In this game, the members of R all make equally essential contributions to earning the unit of payoff, whereas the other players have nothing to contribute to any coalition. Thus, by the same argument that led us to predict the (50,50) allocation in the Divide the Dollars game, the reasonable outcome of this game is to divide the available utility equally among the members of R, giving nothing to the dummy players outside of R; that is,

$$(3) \qquad \phi_i(w_R) = \begin{cases} 1/|R| & \text{if } i \text{ is in } R, \\ 0 & \text{if } i \text{ is not in } R. \end{cases}$$

Using basic results of linear algebra it is straightforward to show that there is a unique value function $\phi = (\phi_1,\ldots,\phi_n)$ that satisfies (1), (2), and (3) above, and this is the Shapley value. Thus, the Shapley value can be understood as the natural linear extension of the equitable solution concept that we applied in Divide the Dollars game. (Nonsymmetric values have been studied by Kalai and Samet [1984].)

A variety of other solution concepts have been defined for cooperative games in characteristic function form. Broadly speaking, these solution concepts can be divided into two categories: equitable solutions and unobjectionable solutions. The unobjectionable solutions include the core, the stable sets or solutions of von Neumann and Morgenstern [1944], the bargaining set (Aumann and Maschler [1964]), and aspiration levels (Bennett [1983]). (See Lucas [1972], Shubik [1982], and Owen [1982] for more about these concepts.) Each of these "unobjectionable" solution concepts tries to identify the set of payoff allocations that are stable, in some sense, against

each coalition's ability to make a demand for some reallocations that it can enforce. The equitable solution concepts include the Shapley value, the Nash bargaining solution, and the nucleolus (Schmeidler [1969]). These "equitable" solution concepts try to identify a reasonable compromise or equitable balance between the various players and coalitions, so that each player's gains from cooperation should be commensurate (in some sense) with what his cooperation contributes to other players.

The simplest way to distinguish between these two categories is by considering the two-player Divide the Dollars game. Each unobjectionable concept identifies the entire set of efficient allocations, from (100,0) to (0,100), as the set of solutions. Every equitable concept identifies the equal division (50,50) as the unique solution to this game.

In Section 6 we discussed Nash's [1951] argument that cooperative games should be analyzed by computing the equilibria of a fully specified model of the bargaining process. As a criticism of the existing literature in cooperative game theory, this argument is more relevant to the unobjectionable solution concepts than to the equitable solution concepts. The unobjectionable solutions are supposed to include all the payoff allocations that the players would accept without forming coalitions to demand reallocation, so it does seem reasonable to ask for a full description of the strategic process by which players form coalitions and make such demands. (Whatever this process is, it always has some equilibria, by the general existence theorem; so the core cannot be identified with the set of equilibria of a bargaining process.)

On the other hand, the equitable solutions can be defended against Nash's argument. As with the Nash bargaining solution, we can interpret the Shapley value and other equitable solution concepts as arbitration guidelines and as determinants of focal equilibria in unarbitrated bargaining, when all

coalitions can negotiate effectively. We only need to assume that the unspecified bargaining process has a sufficiently large set of equilibria, and that the focal equilibrium will be determined by its properties of efficiency and equity, which can be computed from the characteristic function.

The core of a game tends to be much more sensitive to changes in the worths of some coalitions (typically those with larger worths) than others. That is, the core and other unobjectionable solution concepts implicitly identify some coalitions as being more important than others. On the other hand, all worths of coalitions with the same number of members enter into the Shapley-value formula with the same coefficient. Thus, the Shapley value has been criticized for failing to account for the possibility that some coalitions might be more active and important in bargaining than other coalitions of similar size. For example, coalitions that can object to the value might be more active than those with no objections. Some players might have an incentive to make restrictive covenants that prevent them from negotiating separately with other players. For some exogenous reaons (cultural identity, perhaps), some players might be better able to negotiate effectively with each other than with others. To account for such factors, Owen [1977] and Hart and Kurz [1983] have generalized the Shapley value to games with an additional <u>cooperation structure</u> that specifies which coalitions are more active than others.

In the Owen-Hart-Kurz theory, the given cooperation structure consists of a list of certain active coalitions or <u>unions</u>. We let $S_k(i)$ denote the kth largest active union to which player i belongs, and we assume that

$$N \supset S_1(i) \supset S_2(i) \supset \ldots \supseteq \{i\},$$

for every player i. That is, we assume that the unions are <u>nested</u>. If player

j is in $S_k(i)$ then $S_\ell(j) = S_\ell(i)$ for every $\ell \leq k$, since the k largest unions that contain i must also contain j.

Given such a cooperation structure, let us define a permissible ordering of the players to be any strict ordering such that, for any three players h, i, and j, if there is some number k such that $S_k(i) = S_k(j) \neq S_k(h)$, then either i and j both come before h in the ordering or i and j both come after h in the ordering. (That is, a player cannot come in between two members of a union to which he does not belong.) Now suppose that we will randomly select among all the permissible orderings, so that each permissible ordering has equal probability of being selected. Let $\tilde{Q}(i)$ denote the random set of players who come before player i in this randomly selected ordering. The Owen-Hart-Kurz (OHK) value for player i, in a game with characteristic function v and the given cooperation structure (denoted by $(S_1(\cdot),S_2(\cdot),\ldots)$), is defined to be

$$\phi_i(v|S_1,S_2,\ldots) = E(v(\tilde{Q}(i) \cup \{i\}) - v(\tilde{Q}(i)))$$

(where E denotes the expected value). So the OHK value for i is his expected marginal contribution to the random coalition that precedes him.

If $S_1(i) = \{i\}$ for every player i, so that there are no multi-player unions, then every ordering of the players is permissible. In this case, the OHK value is equal to the Shapley value.

With the OHK value as our solution concept, we may try to analyze games to predict which unions of players are most likely to become active. In the three-player majority game (Example 13, above), players 1 and 2 can gain by forming the union $\{1,2\}$, since it would increase each of their values from 100 (the Shapley value) to 150. On the other hand, in the three-player unanimity game (Example 11, above), players 1 and 2 would lose by forming the union

$\{1,2\}$, since it would decrease each of their payoffs from 100 to 75. In general, the development of a plausible model of endogenous determination of cooperation structures remains an important unsolved problem in game theory. A theory of bargaining that is based on a value for games with cooperation structure and on a plausible model of endogenous union formation, could combine the best properties of the equitable and unobjectionable solution theories.

8. Cooperative Games Without Transferable Utility

To extend the Shapley value (and other solution concepts similarly) to games without transferable utility, Shapley [1969] suggested the following "λ-transfer" theory. Given a strategic-form game

$$\Gamma = (C_1,\ldots,C_n,\ u_1,\ldots,u_n)$$

as in Section 2, and given any vector $\lambda = (\lambda_1,\ldots,\lambda_n)$ such that all $\lambda_i > 0$, let the <u>λ-rescaled version</u> of Γ be

$$\lambda^*\Gamma = (C_1,\ldots,C_n,\ \lambda_1 u_1,\ldots,\lambda_n u_n).$$

That is, $\lambda^*\Gamma$ differs from Γ only in that the utility function of each player i is multiplied by λ_i. Without transferable utility, there is no decision-theoretically testable distinction between these two games. So let us consider any such rescaled version $\lambda^*\Gamma$ and analyze it as if the λ-weighted utilities were freely transferable, computing its characteristic function v^λ and its Shapley value $\phi(v^\lambda) = (\phi_1(v^\lambda),\ldots,\phi_n(v^\lambda))$. Let x_i^λ be the payoff for player i in the original utility scales of Γ that corresponds to the payoff $\phi_i(v^\lambda)$ in the λ-weighted utility scales of $\lambda^*\Gamma$; that is

$$x_i^\lambda = \frac{1}{\lambda_i}\,\phi_i(v^\lambda).$$

In general, the allocation $x^\lambda = (x_1^\lambda,\ldots,x_n^\lambda)$ would be feasible in Γ if λ-weighted utility were transferable; but x^λ is usually not feasible without such transfers. However, if x^λ actually is feasible in Γ, without any transfers of utility, then we say that x^λ is a Shapley NTU value (or a λ-transfer value) for Γ, and λ is a vector of natural utility weights for Γ. (Here NTU stands for "nontransferable utility.") The existence of a Shapley NTU value can be guaranteed (see Shapley [1969] and Myerson [1984b]) if we allow any vector λ in which all components are nonnegative. (For vectors in which some λ_i are zero, we may define x^λ as any limit of a sequence of $x^{\lambda(k)}$ allocations, such that all components of each vector $\lambda(k)$ are positive and the vectors $\lambda(k)$ converge to λ as k goes to infinity.)

Alternative definitions of NTU values have been suggested by Harsanyi [1963] and Owen [1972]. Axiomatic derivations of the Shapley NTU value and the Harsanyi NTU value have recently been developed by Aumann [1983] and Hart [1983]. (See also Samet [1984].) In the case of games with two players, all three of these NTU values are equal to the Nash bargaining solution, with the same natural utility weights as in Section 6. For games with transferable utility, these three NTU values all equal the Shapley value.

Roth [1980] and Shafer [1980] have shown examples in which the Shapley NTU value selects outcomes that seem intuitively to be very unreasonable. The Harsanyi NTU value seems somewhat more reasonable for Roth's examples. (The Owen NTU value seems too complicated to compute.) On the other hand, Myerson [1984b] has been able to define a natural extension of the Shapley NTU value to games with incomplete information, but not the more complicated and nonlinear Harsanyi NTU value. Thus, the Shapley NTU value stands as the most

broadly defined natural extension of our two most compelling solution concepts: the Shapley value (for games with transferable utility) and the Nash bargaining solution (for games with two players). Furthermore, there is reason to hope that some modification of the Shapley NTU value, perhaps based on the OHK value with endogenously determined cooperation structures, could provide a satisfactory analysis of all examples. Thus it is important to try to understand the logic behind the Shapley NTU value.

Consider the Banker Game from Owen [1972]. In this three-player game, the coalition $\{1,2\}$ can achieve any nonnegative utility allocation (y_1,y_2) such that $y_1 + 4y_2 \leq 100$. The grand coalition $\{1,2,3\}$ can achieve any nonnegative utility allocation (y_1,y_2,y_3) such that $y_1 + y_2 + y_3 \leq 100$. Every other coalition can only get zero for its members. The idea is that player 1 can get \$100 with the help of player 2. To reward player 2 for his help, player 1 can try to send him money; but without player 3, there is a 75% chance of losing the money that is sent. Player 3 is a banker who can prevent such loss in transactions. How much should player 1 pay to player 2 for his help and to player 3 for his banking services?

The unique Shapley NTU value for this game is (50,50,0), supported by the natural utility weights $\lambda = (1,1,1)$. With these weights, $v^\lambda(\{1,2\}) = 100$, because the maximum λ-weighted sum of utilities that coalition $\{1,2\}$ can get is 100, at $(y_1,y_2) = (100,0)$. Also, $v^\lambda(\{1,2,3\}) = 100$, and every other coalition S gets $v^\lambda(S) = 0$. The Shapley value of this v^λ is (50,50,0).

Owen [1972] argued that player 1 should get more than player 2, and that player 3 should get some positive fee for his banking services; but there is a rationale to this Shapley NTU value. Getting zero, player 3 is indifferent between accepting this NTU-value outcome or not, so it is not unreasonable to assume that he probably will accept it. (Think of his NTU-value payoff as

positive but infinitesimal, while his cost of providing banking services is zero.) So suppose that there is only some small probability q that player 3 will refuse to accept his NTU-value allocation and will break up the grand coalition. As long as $q \leq 1/2$, players 1 and 2 can accommodate this possibility with no loss of expected utility. They simply plan to choose (100,0) if 3 rejects the grand coalition (no transfer of money without the banker), and plan to choose $(100 - 50/(1 - q), 50/(1 - q), 0)$ if 3 agrees to cooperate (a transfer of $50/(1 - q)$ using the banker).

Now let i equal 1 or 2; and suppose instead that there were a small probability q that player i would reject the NTU-value outcome, even though it is better for him than the zero that he gets alone. In this case, the expected payoffs to the other two players could not sum to more than $50(1 - q)$ without reducing player i's allocation in the case of agreement. Thus, a low-probability threat of rejection by either player 1 or 2 would cause real losses in the expected payoffs of the other players, and in a symmetrical manner; but such a threat by player 3 would have no effect on expected payoffs if it were anticipated correctly. In this sense, players 1 and 2 have equal power and player 3 has none, so that (50,50,0) is a reasonable bargaining solution.

In general, let x be an efficient payoff allocation for the grand coalition in a given game. Let λ be a vector of utility weights such that x maximizes the sum of λ-weighted utilities $(\lambda_1 x_1 + \ldots + \lambda_n x_n)$ over all payoff allocations that are feasible for the grand coalition. Suppose that the efficient frontier is differentiable or smooth at x. Then, to a first-order approximation, small transfers of λ-weighted utility are feasible near x for players in the grand coalition. That is, for any sufficiently small δ, if player i reduced his utility payoff from x_i to $x_i - \delta/\lambda_i$ (sacrificing δ

units of λ-weighted utility) then, without changing any other players' payoffs from what they get in the allocation x, player j could increase his utility payoff from x_j to $x_j + \delta/\lambda_j$, minus some "transactions cost" that is small in proportion to δ.

Now suppose that the players are expected to unanimously accept the allocation x almost surely, except that, with some small probability, a smaller coalition S might have to choose something feasible for themselves. In this situation, a small transfer of λ-weighted utility in the event that everyone accepts x would have the same effect on expected payoffs as a large transfer of λ-weighted utility in the event that coalition S acts alone. Thus, when the members of coalition S plan what to do if they must act alone, they can effectively transfer λ-weighted utility among themselves, where the coin of transfer is a promise to make a small feasible reallocation around x in the much more likely event that x is accepted. (The players outside S would not object to such reallocation because it does not affect their payoffs. We are assuming that these coalitional plans are made before it is learned whether the coalition must act alone or not.) So it is appropriate to analyze this bargaining game as if λ-weighted utility really were transferable for any such coalition S. The results of this analysis (when we compute the Shapley value of the λ-rescaled version of the game, and then convert this value back into the original utility scales) will coincide with the originally hypothesized allocation x if and only if x is a Shapley NTU value. In this sense, the Shapley NTU values are the plausible cooperative solutions of the game.

9. Cooperative Games with Incomplete Information

In a cooperative game with incomplete information, the players already know their private information or "type" when they bargain over which communication mechanism to implement. Recall that a communication mechanism for a Bayesian game with incomplete information is a rule for determining the actions of all players as a (possibly random) function of reports that the players submit to some mediator. Let us suppose that the players can make binding commitments to regulate their actions but cannot verifiably audit each others' types. Thus, a communication mechanism μ is feasible for the players together only if it satisfies the informational incentive constraints discussed near the end of Section 5.

If player i's actual type is t_i, then his objective in bargaining is to maximize his conditionally expected payoff $U_i(\mu|t_i)$ given his actual type. His conditionally expected payoff given any other possible type, and his ex ante expected payoff before his type was learned, would be completely irrelevant to his welfare, since he already knows his actual type. However, an outside arbitrator, who does not know any player's actual type, can be sure that all players would want to make some change in their communication maecahism only if the change increased the conditionally expected payoffs $U_i(\mu|t_i)$ for every type t_i of every player i. From such an outsider's viewpoint, if there are three players and if there are five possible types or information states for each player, then a change is an unambiguous welfare improvement only if it increases (or at least does not decrease) each of the fifteen conditionally expected payoffs for the various possible types of the players. In general, we may say a communication mechanism μ is efficient if it is feasible and there does not exist any other feasible communication mechanism ν such that

$$U_i(\nu|t_i) \geq U_i(\mu|t_i)$$

for every possible type t_i of every player i, with at least one strict inequality. (For a comprehensive discussion of efficiency with incomplete information, see Holmstrom and Myerson [1983]. We are here defining efficiency to mean interim incentive-efficiency, in the terminology of Holmstrom and Myerson.)

In bargaining without an arbitrator, the expected payoffs for all possible types of a player may still be relevant to the bargaining process, because the other players do not know his actual type and he may wish to conceal it. If a player were expected to demand the feasible communication mechanism that maximizes his conditionally expected utility given his actual type, then his demand could reveal his type-information to the other players, and they might be able to use this information against him. Thus, a player's optimal bargaining strategy should represent some kind of inscrutable compromise between his actual preferences and the preferences that he would have had if his informational type had been different. Therefore, a cooperative bargaining solution should be an equitable compromise, in some sense, not only between all the different players, but also between all the different possible types of each player.

Based on such considerations of efficiency and equity, Myerson [1983, 1984a, 1984b] has defined neutral bargaining solutions, which generalize the Nash bargaining solution and the Shapley NTU value to games with incomplete information. These neutral bargaining solutions satisfy equity and efficiency properties that can be described in terms of certain virtual-utility functions. Without giving a formal definition here, we may say that a player's virtual utility differs from his real utility by taking into account the costs of satisfying his informational incentive constraints. (In a sense,

the definition of virtual utility is an application of one of the most basic ideas of economic theory: that efficient social plans could be decentralized if the constraints facing society were multiplied by some appropriate shadow prices and added into the individual's payoff functions. The only difference is that here we are considering incentive constraints, instead of resource constraints. For a full basic explanation of virtual utility see Myerson [1985a].) The essential idea of these neutral bargaining solutions is to apply the Shapley value (in each information state) to a transformed game in which the players get transferable virtual-utility payoffs, in the same way that the Shapley NTU value applies the Shapley value to a transformed game in which the players get transferable weighted-utility payoffs. If the resulting allocation of virtual utility corresponds to an allocation of real utility that can actually be achieved by a feasible communication mechanism then that mechanism is a neutral bargaining solution.

It is best to introduce these ideas in the context of a simple two-player example. Let player 1 be a (monopolistic) seller and let player 2 be a (monopsonistic) buyer of some commodity. The seller has a supply of one unit of the commodity, and he knows whether it is good quality (type "1a") or bad quality (type "1b"). If it is good quality then it is worth \$40 per unit to the seller and \$50 per unit to the buyer. If it is bad quality then it is worth \$20 per unit to the seller and \$30 per unit to the buyer. The buyer thinks that the probability of good quality is 0.2. We assume that the buyer cannot verifiably audit the seller's information, and the seller cannot offer any enforceable warranties. They must simply negotiate a price and quantity to be traded, possibly depending on what the seller claims about his information.

To describe a <u>trading mechanism</u> that the players might use, let x_a and q_a

denote, respectively, the amount of money that the buyer will pay to the seller and the quantity of the commodity that the seller will give to the buyer if the seller claims that the quality is good; and let x_b and q_b denote corresponding quantities if the seller claims that the quality is bad. If the seller is honest in such a trading mechanism then his expected payoff is

$$U_{1a} = x_a - 40q_a \quad \text{if the commodity is good, and}$$

$$U_{1b} = x_b - 20q_b \quad \text{if the commodity is bad.}$$

The buyer, who does not know the quality, gets the expected payoff

$$U_2 = (.2)(50q_a - x_a) + (.8)(30q_b - x_b).$$

To be feasible, a trading mechanism must satisfy the following two informational incentive constraints

$$U_{1a} \geqslant x_b - 40q_b, \quad U_{1b} \geqslant x_a - 20q_a,$$

so that the seller cannot gain by lying about his information. Also, since each player has the option to not trade at all (which gives him a payoff of zero) a feasible trading mechanism must also satisfy the following three <u>minimum-payoff contraints</u> (often called <u>individual-rationality</u> constraints)

$$U_{1a} \geqslant 0, \quad U_{1b} \geqslant 0, \quad U_2 \geqslant 0.$$

In addition, we must have $0 \leqslant q_a \leqslant 1$ and $0 \leqslant q_b \leqslant 1$, since there is only one unit to trade.

Notice that the commodity is always worth \$10 more to the buyer than to the seller. However, there is no feasible trading mechanism in which the buyer always gets all of the seller's supply. Such a mechanism would have $q_a = q_b = 1$; but then the incentive constraints would imply that $x_a = x_b$,

so that either $U_{1a} < 0$ or $U_2 < 0$. Thus, by the revelation principle (see Section 5), in any equilibrium of any bargaining process applied to this game, there must be some positive probability that the seller will end up owning some of the commodity, even though it should be worth more to the buyer. The problem is that the good-type seller (1a) cannot convincingly demonstrate that he really needs and deserves a price above \$40, unless he implements a threat to withhold some of his supply. Without such a demonstration, the buyer would be unwilling to pay more than \$34 = (.8)(30) + (.2)(50).

There are many trading mechanisms that do satisfy all of the feasibility constraints, however. Analysis of these constraints shows that, for any numbers q_a and q_b such that $0 \leq q_a \leq (4/7)q_b$ and $q_b \leq 1$, there exist some x_a and x_b that make a feasible trading mechanism. (See Proposition 3 in Myerson [1985b].) In general, the good-type seller always sells strictly less than the bad type, but at a higher price per unit.

To determine which trading mechanisms are efficient, we must characterize the set of all allocations of expected utility (U_{1a},U_{1b},U_2) to the two types of seller and the buyer that can be achieved using a feasible trading mechanism. It can be shown, by mathematical analysis of the feasibility constraints, that an allocation (U_{1a},U_{1b},U_2) can be achieved if and only if it satisfies the following five inequalities:

$$.3U_{1a} + .7U_{1b} + U_2 \leq 8, \quad U_{1a} \leq U_{1b}, \quad U_{1a} \geq 0, \quad U_{1b} \geq 0, \quad \text{and} \quad U_2 \geq 0.$$

Thus, any feasible meachism that satisfies

$$.3U_{1a} + .7U_{1b} + U_2 = 8$$

must be efficient, in the sense that there is no other feasible trading

mechanism that would be surely preferred by the seller (in either type) and by the buyer.

The best feasible mechanism for the buyer is

(4) $$q_a = 0, \quad x_a = 0, \quad q_b = 1, \quad x_b = 20$$

which gives expected payoffs

$$U_{1a} = 0, \quad U_{1b} = 0, \quad U_2 = 8.$$

This mechanism is implemented by letting the buyer make a nonnegotiable first-and-final offer to buy the seller's unit of supply for \$20, accepting the 20% chance that the seller might refuse to trade because he is a good type. To increase q_a above zero, it would be necessary to offer a higher price to the bad-type seller, which would reduce the buyer's expected payoff.

The best feasible mechanism for the seller depends on his type. For the good type (1a) the best feasible mechanism is

(5) $$q_a = 0, \quad x_a = 8, \quad q_b = 1, \quad x_b = 28$$

which gives

$$U_{1a} = 8, \quad U_{1b} = 8, \quad U_2 = 0.$$

This mechanism differs from the buyer's best (4) in that the buyer first has to pay a nonrefundable fee of \$8, to buy the right to then make a final offer of \$20 for the commodity. The good-type seller would take the \$8 fee and then refuse to sell for \$20. In expected value, the buyer is compensated for his potential losses to the good type by his \$2 gains from the more likely bad type. On the other hand, the best feasible mechanism for the bad type (1b) is

(6) $$q_a = \frac{4}{7}, \quad x_a = 22\frac{6}{7}, \quad q_b = 1, \quad x_b = 31\frac{3}{7},$$

which gives

$$U_{1a} = 0, \quad U_{1b} = 11\tfrac{3}{7}, \quad U_2 = 0.$$

In this mechanism, the buyer buys 4/7 units from the good type at a price per unit of $x_a/q_a = 40$. So his gains from the good type compensate, in expected value, for his losses when he pays a price above \$30 to the bad type.

Under the assumption that the buyer and seller have equal bargaining ability, the neutral bargaining solution selects the trading mechanism

$$(7) \qquad q_a = 1/6, \quad x_a = 50/6, \quad q_b = 1, \quad x_b = 25$$

which gives

$$U_{1a} = 1\tfrac{2}{3}, \quad U_{1b} = 5, \quad U_2 = 4.$$

(Notice that $.3(10/6) + .7(5) + 4 = 8$, so this is efficient.) In this mechanism, the seller can either sell his entire supply for \$25, or he can sell 1/6 of his supply for a price per unit of \$50 = x_a/q_a; the bad type chooses the former and the good type chooses the latter. The price of \$25 seems clearly equitable for the bad type of seller (averaging the seller's valuation of \$20 and the buyer's valuation of \$30), but the \$50 price for the good type fully exploits the buyer. However, it can be shown (see Myerson [1985a]) that if the costs of the incentive constraints were internalized using the hypothetical construction of virtual utility, the good-type seller's virtual valuation would become \$50 instead of \$40, so the \$50 price satisfies the property of "virtual equity" for the good type. The intuitive idea behind the virtual-utility hypothesis is that the good type of seller is jeopardized by the bad type (that is, type 1a needs to prove to the buyer that he is not type 1b), so that the good type might tend to distort his effective

preferences and act as if the commodity was worth \$50 to him rather than \$40, to exaggerate his difference from the bad type. These exaggerated virtual preferences could also justify the outcome that only a fraction of the good seller's supply is sold, since his virtual valuation equals the buyer's valuation for the commodity.

To understand the rationale behind the neutral bargaining solution more rigorously, it is necesssary to face the issue of inscrutable compromise between two types of the seller. The buyer's expected payoff $U_2 = 4$ seems equitable, in that it is halfway between the best $(U_2 = 8)$ and worst $(U_2 = 0)$ that he could expect in any feasible trading mechanism. There are many possible allocations for the different types of seller (from $(U_{1a},U_{1b}) = (4,4)$ to $(U_{1a},U_{1b}) = (0,5\frac{5}{7})$) that are all achievable with feasible trading mechanisms in which the buyer's expected payoff is 4, but only if both types of seller are expected to use the same mechanism (so that the choice of mechanism does not alter the buyer's beliefs). What is special about the allocation $(U_{1a},U_{1b}) = (1\frac{2}{3},5)$ that makes it the most reasonable or inscrutable compromise between the conflicting interests of the two types of seller (so that the buyer should not infer anything about the seller's type from the fact that he is willing to settle for a mechanism that gives this allocation)?

Consider first the simpler case in which the seller has all of the bargaining ability (or the seller is a <u>principal</u> in the sense discussed at the end of Section 3). In this case, the seller does not need to compromise with the buyer, who will presumably accept any trading mechanism that gives him a nonnegative expected payoff. However, the seller must still make some compromise between his actual type and the other possible type, to avoid conveying information to the buyer by the mechanism selection itself. That

is, the seller cannot simply demand the feasible mechanism that he likes best given his actual type, because the buyer would reject mechanism (5) when he realizes that only the good type would want to implement it, and the buyer would reject (6) when he realizes that only the bad type would implement it. (The buyer would expect to lose \$8 to the good type in (5) and lose $\$1\frac{3}{7}$ to the bad type in (6).) The most inscrutable compromise for the informed seller would be

$$(8) \qquad q_a = 1/3, \quad x_a = 50/3, \quad q_b = 1, \quad x_b = 30.$$

In this feasible mechanism, the bad type sells his entire supply for \$30 and the good type sells 1/3 of his supply for \$50 per unit. The buyer's payoff is zero with either type of seller, so the buyer would be willing to participate in this mechanism no matter what he inferred about the seller's type from the fact that the seller proposed it. Furthermore, this mechanism is efficient and gives $U_2 = 0$, so there is no feasible mechanism that makes both types of the seller better off. These properties make mechanism (8) a <u>strong optimum</u> for the seller, in the sense of Myerson [1983]. It can be shown that, for any alternative feasible mechanism that is better for one type of the seller, the buyer would expect negative payoff in this alternative mechanism if he inferred that the seller's type is the one that prefers it. So any other proposal by the seller would be rejected by the buyer, on the basis of the information revealed by the proposal itself.

The expected payoffs from the seller's optimum (8) are

$$U_{1a} = 3\tfrac{1}{3}, \quad U_{1b} = 10, \quad U_2 = 0.$$

The averages of these payoffs with those of the buyer's optimum (4) are exactly equal to the expected payoffs of the neutral bargaining solution

(7). That is, the neutral bargaining solution is equivalent to a randomization between the buyer's optimum (4) and the seller's optimum (8), in which each mechanism gets equal probability. This random-dictatorship property is in fact one of the two axioms (the other being an analogue of Nash's axiom of independence of irrelevant alternatives) from which the neutral bargaining solution was first derived by Myerson [1984a].

It can also be instructive to analyze this game by the noncooperative approach to bargaining, characterizing the equilibria of a specific bargaining process. Similar games have been analyzed in this way by many authors, including Fudenberg and Tirole [1983], Cramton [1983], Sobel and Takahashi [1983], Rubinstein [1983], and Chatterjee and Samuelson [1983]. By the revelation principle, any equilibrium of any such bargaining process will be equivalent to some feasible mechanism as described above. Unfortunately, many natural bargaining processes turn out to have multiple equilibria.

For example, consider the bargaining process in which the seller first sets a price per unit, and then the buyer decides what quantity to purchase. There are infinitely many sequential equilibria of this game. For any price y between 40 and 50, there is a sequential equilibrium in which the good type of the seller sets a price of y. The bad type randomizes between setting a price of 30, with probability $(5y - 170)/(4y - 120)$, and a price of y, with probability $(50 - y)/(4y - 120)$. When the bad type randomizes in this way, the buyer would rationally believe, after getting a price of y, that the commodity should be worth \$y per unit to him, because

$$y = \frac{(30)(.8)(50 - y)/(4y - 120) + (50)(.2)}{(.8)(50 - y)/(4y - 120) + .2}.$$

So the buyer's demand can be rationally set at $10/(y - 30)$ units of the

commodity after getting a price of y, and at one unit after getting a price of 30. For any price above 30 other than y, we may suppose that the buyer would demand no units of the commodity, because he might infer that the unexpected price quote came from the bad type of seller. Notice that the bad type of seller is indifferent between setting the price at 30 or at y, as is necessary to induce him to randomize.

All of these equilibria correspond to efficient trading mechanims, and all give the same expected payoff to the bad type of seller. The difference is that the buyer prefers the equilibria with lower y and the good type of seller prefers the equilibria with higher y. (There are other, inefficient equilibria of this game which are worse for both the buyer and the good type of seller than equilibria described above.) Thus, if we add the assumption that the seller not only sets the price but also has the persuasive power of a principal to determine the equilibrium (that is, he can explain which sequential equilibrium he is implementing when he sets his price, and the buyer will accept his explanation), then we should expect that the sequential equilibrium with y = 50 will be implemented. This equilibrium is equivalent to the seller's neutral optimum (trading mechanism (7)) which we discussed above.

References

R. J. Aumann [1974], "Subjectivity and Correlation in Randomized Strategies," Journal of Mathematical Economics 1, 67-96.

R. J. Aumann [1976], "Agreeing to Disagree," Annals of Statistics 4, 1236-1239.

R. J. Aumann [1981], "Survey of Repeated Games," in Essays in Game Theory and Mathematical Economics, by R. J. Aumann et al., Zurich: Bibliographisches Institut, 11-42.

R. J. Aumann [1983], "An Axiomatization of the Non-transferable Utility Value," RM-57, Center for Research in Mathematical Economics and Game Theory, The Hebrew University, Jerusalem.

R. J. Aumann and M. Maschler [1964], "The Bargaining Set for Cooperation Games," in Advances in Game Theory, ed. by M. Dresher, L. S. Shapley, and A. W. Tucker, Princeton: Princeton University Press, 443-447.

R. J. Aumann and M. Maschler [1972], "Some Thoughts on the Minimax Principle," Management Science 18, P54-P63.

R. J. Aumann and L. S. Shapley [1974], Values of Non-Atomic Games, Princeton: Princeton University Press.

E. Bennett [1983], "The Aspiration Approach to Predicting Coalition Formation and Payoff Distribution in Sidepayment Games," International Journal of Game Theory 12, 1-28.

B. D. Bernheim [1984], "Rationalizable Strategic Behavior," Econometrica 52, 1007-1028.

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

K. Chatterjee and W. Samuelson [1983], "Bargaining Under Incomplete Information," Operations Research 31, 835-851.

P. Cramton [1984], "The Role of Time and Information in Bargaining," discussion paper, Graduate School of Business, Stanford University.

V. Crawford [1983], "Efficient and Durable Decision Rules: A Reformulation," discussion paper, Department of Economics, University of California, San Diego.

E. van Damme [1984], "A Relation Between Perfect Equilibria in Extensive Games and Proper Equilibria in Normal Form Games," International Journal of Game Theory 13, 1-13.

D. Fudenberg and J. Tirole [1983], "Sequential Bargaining with Incomplete Information," Review of Economic Studies 50, 221-247.

J. C. Harsanyi [1983], "A Simplified Bargaining Model for the n-Person Cooperative Game," International Economic Review 4, 194-220.

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 [1973], "Games with Randomly Disturbed Payoffs: A New Rationale for Mixed-Strategy Equilibrium Points," International Journal of Game Theory 2, 1-23.

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.

J. C. Harsanyi and R. Selten [1985], A General Theory of Equilibrium Selection in Games, forthcoming. (Available in discussion papers of the Institute for Mathematical Economics of the University of Bielefeld.)

S. Hart [1983], "An Axiomatization of Harsanyi's Non-transferable Utility Solution," discussion paper, the Center for Mathematical Studies in Economics and Management Sciencs, Northwestern University.

S. Hart and M. Kurz [1983], "Endogenous Formation of Coalitions," Econometrica 51, 1799-1819.

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

K. Kalai [1977], "Nonsymmetric Nash Solutions and Replications of Two-Person Bargaining," International Journal of Game Theory 6, 129-133.

E. Kalai and D. Samet [1982], "Persistent Equilibria in Strategic Games," discussion paper, Northwestern University.

E. Kalai and D. Samet [1984], "On Weighted Shapley Values," discussion paper, Northwestern University.

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

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

H. W. Kuhn [1953], "Extensive Games and Problems of Information," in Contributions to The Theory of Games II, edited by H. W. Kuhn and A. W. Tucker, Princeton: Princeton University Press, 193-216.

W. F. Lucas [1972], "An Overview of the Mathematical Theory of Games," Management Science 18, P3-P19.

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

J. R. Mertens and S. Zamir [1983], "Formalization of Harsanyi's Notion of 'Type' and 'Consistency' in Games with Incomplete Information," CORE discussion paper, Universite Catholique de Louvain.

P. R. Milgrom and R. J. Weber [1984], "Distributional Strategies for Games with Incomplete Information," discussion paper, Northwestern University, to appear in Mathematics of Operations Research.

R. B. Myerson [1978a], "Refinements of the Nash Equilibrium Concept," International Journal of Game Theory 7, 73-80.

R. B. Myerson [1978b], "Threat Equilibria and Fair Settlements in Cooperative Games," Mathematics of Operations Research 3, 265-274.

R. B. Myerson [1979], "An Axiomatic Derivation of Subjective Probability, Utility, and Evaluation Functions," Theory and Decision 11, 339-352.

R. B. Myerson [1982], "Optimal Coordination Mechanisms in Generalized Principal-Agent Problems," Journal of Mathematical Economics 10, 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], "Acceptable and Predominant Correlated Equilibria," discussion paper, Northwestern University, to appear in International Journal of Game Theory.

R. B. Myerson [1984d], "Multistage Games with Communication," discussion paper, Northwestern University, to appear in Econometrica.

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.

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

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, 128-140.

J. von Neumann and O. Morgenstern [1944], Theory of Games and Economic Behavior, Princeton: Princeton University Press.

G. Owen [1972], "Values of Games Without Sidepayments," International Journal of Game Theory 1, 94-109.

G. Owen [1977], "Values of Games with A Priori Unions," in Essays in Mathematical Economics and Game Theory, edited by R. Hein and O. Moeschlin, Berlin: Springer-Verlag, 76-88.

G. Owen [1982], Game Theory (2nd edition), New York: Academic Press.

D. G. Pearce [1984], "Rationalizable Strategic Behavior and the Problem of Perfection," Econometrica 52, 1029-1050.

H. Raiffa [1968], Decision Analysis, Reading, Massachusetts: Addison-Wesley.

A. E. Roth [1979], Axiomatic Models of Bargaining, Berlin: Springer-Verlag.

A. E. Roth [1980], "Values for Games Without Side Payments: Some Difficulties with Current Concepts," Econometrica 48, 457-465.

A. E. Roth and F. Schoumaker [1983], "Expectations and Reputations in Bargaining," American Economic Review 73, 362-372.

A. Rubinstein [1979], "Equilibrium in Supergames with the Overtaking Criterion," Journal of Economic Theory 21, 1-9.

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

D. Samet [1984], "An Axiomatization of the Egalitarian Solutions," to appear in Mathematical Social Sciences.

L. J. Savage [1954], The Foundations of Statistics, New York: John Wiley and Sons.

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

T. C. Schelling [1960], The Strategy of Conflict, London: Oxford University Press.

D. Schmeidler [1969], "The Nucleolus of a Characteristic Function Game," SIAM Journal of Applied Mathematics 17, 1163-1170.

W. Shafer [1980], "On the Existence and Interpretation of Value Allocations," Econometrica 48, 467-477.

L. S. Shapley [1953], "A Value for n-Person Games," in Contributions to the Theory of Games 2, edited by H. Kuhn and A. W. Tucker, Princeton: Princeton University Press, 307-317.

L. S. Shapley [1969], "Utility Comparison and the Theory of Games," in La Decision, Paris: Editions du CNRS, 251-263.

M. Shubik [1982], Game Theory in the Social Sciences, Cambridge: MIT Press.

J. Sobel and I. Takahashi [1983], "A Multistage Model of Bargaining," Review of Ecomomic Studies 50, 411-426.