Incentive Compatibility and the Bargaining Problem

Author(s): Roger B. Myerson

Source: *Econometrica*, Jan., 1979, Vol. 47, No. 1 (Jan., 1979), pp. 61-73

Published by: The Econometric Society

Stable URL: https://www.jstor.org/stable/1912346

Econometrica, Vol. 47, No. 1 (January, 1979)

INCENTIVE COMPATIBILITY AND THE BARGAINING PROBLEM

BY ROGER B. MYERSON

Collective choice problems are studied from the Bayesian viewpoint. It is shown that the set of expected utility allocations which are feasible with incentive-compatible mechanisms is compact and convex, and includes the equilibrium allocations for all other mechanisms. The generalized Nash solution proposed by Harsanyi and Selten is then applied to this set to define a bargaining solution for Bayesian collective choice problems.

1. INTRODUCTION

IN THIS PAPER we will consider the problem of an arbitrator trying to select a collective choice for a group of individuals when he does not have complete information about their preferences and endowments. Not only does this arbitrator have to worry about settling fairly the conflicting desires of the group's diverse members, but he has to get them to tell him what their preferences are in the first place. Of course, he may ask his clients to tell him what he needs to know; but if he cannot compel truthful behavior then he must anticipate that some group members may lie to him in an attempt to manipulate his ultimate decision. Our goal in this paper will be to develop a unique solution to this arbitrator's problem, based on Hurwicz's concept of *incentive-compatibility* [**6**] and Nash's *bargaining solution* [**7**].

Formally, we will describe the arbitrator's problem by a *Bayesian collective choice problem*, an object of the form

$$(C, A_1, A_2, \ldots, A_n, U_1, U_2, \ldots, U_n, P), \tag{1}$$

whose components are interpreted as follows. The individual members of the group, or *players*, are numbered $1, 2, \ldots, n$. C is the set of choices or strategies available to the group. For each player i, A_i is the set of possible *types* for player i. That is, each $a_i \in A_i$ represents a complete description of player i's relevant characteristics: his preferences, beliefs, abilities, and endowments. Each U_i is a utility function from $C \times A_1 \times \ldots \times A_n$ into the real numbers such that each $U_i(c, a_1, a_2, \ldots, a_n)$ is the payoff which player i would get if $c \in C$ were chosen and if $(a_1, \ldots, a_n)$ were the true vector of player types. These U_i payoff numbers are assumed to be measured in terms of some von Neumann-Morgenstern utility scale for player i. Finally, P is a probability distribution on $A_1 \times \ldots \times A_n$ such that $P(a_1, \ldots, a_n)$ is the probability, as judged by the arbitrator, that $(a_1, \ldots, a_n)$ is the true vector of types for the n players.

To avoid mathematical complications, we shall assume that C and all the A_i sets are nonempty finite sets. That is, there are only finitely many types which are considered possible for each player, and there are only a finite number of choice options available to the group. However, we will admit randomized strategies (as proposed in [**3**]). That is, instead of selecting a particular choice in C, the arbitrator may select a probability distribution over C and let the actual choice in C be determined randomly according to this distribution.

An arbitrator's solution to a collective choice problem would typically be a procedure in which he first asks every player for some information about his type, and then selects a choice in C, or a probability distribution over C, using the information which the players have given him. To formally model such procedures, we define a *choice mechanism* to be a real-valued function π with a domain of the form $C \times (S_1 \times S_2 \times \ldots \times S_n)$—for some collection of *response sets* $S_1, S_2, \ldots, S_n$—such that

$$\sum_{c' \in C} \pi(c'|s_1, \ldots, s_n) = 1, \quad \text{and} \quad \pi(c|s_1, \ldots, s_n) \geq 0 \quad \text{for all } c, \tag{2}$$

for every

$$(s_1, \ldots, s_n) \quad \text{in} \quad S_1 \times \cdots \times S_n.$$

Here each S_i is interpreted as the set of possible responses which player i might give to the arbitrator's questions; and each number $\pi(c|s_1, \ldots, s_n)$ is interpreted as the probability which the arbitrator will assign to choice c if $(s_1, \ldots, s_n)$ is the combination of responses which he gets from the players.

If the arbitrator simply asks each player what his true type is, then player i's response set should be $S_i = A_i$, since any of his possible types could be a plausible answer. We shall refer to the A_i as the *standard response sets*. Except in this section and in Section 3, we will always restrict our attention to choice mechanisms on the standard response sets.

We shall assume that the response of each player is communicated to the arbitrator confidentially and noncooperatively. Thus, when player i selects his response s_i in S_i, he does not know what any other player's response has been, and he must select his response independently of any other player's decision.

Depending on what the arbitrator asks the players, there may be some *truthful response map* $\tau_i : A_i \to S_i$, such that $\tau_i(a_i)$ would be i's truthful response if he were of type a_i. For the standard response sets, the natural truthful response map is the identity map $\tau_i(a_i) = a_i$. We will *not* assume that the arbitrator has any way to compel a player to give the truthful response. Each player is the only one who can know his own true type for sure, and no one can prevent him from lying about it when he expects advantage from lying. On the other hand, when there is no positive incentive to lie, we may expect the players to be honest.

We shall assume that the players will always accept the arbitrator's final recommended choice in C; that is, the arbitration is binding.

Finally, we shall assume a version of the *consistency condition* of Harsanyi [**4**]. To express this in our framework, we need some definitions. Observe first that each player i is given the same information as the arbitrator (he knows the basic structure of (1)) plus one additional fact: player i also knows his own true type a_i in A_i. If the arbitrator were to learn that player i was of type a_i, then the arbitrator would reassess the posterior (conditional) probability of the types vector $\boldsymbol{\alpha} = \alpha_1, \ldots, \alpha_n)$ to be

$$P_i(\alpha_1, \ldots, \alpha_n | a_i) = \begin{cases} P(\alpha_1, \ldots, \alpha_n)/R_i(a_i) & \text{if} \quad \alpha_i = a_i, \\ 0 & \text{if} \quad \alpha_i \neq a_i, \end{cases} \tag{3}$$

where $R_i(a_i)$ is the marginal probability of a_i,

$$(4) \qquad R_i(a_i) = \sum_{\substack{(\beta_1,\ldots,\beta_n)\in A_1\times\ldots\times A_n \\ \text{such that } \beta_i = a_i}} P(\beta_1, \ldots, \beta_n).$$

Our *consistency* assumption is that $P_i(\alpha_1, \ldots, \alpha_n|a_i)$ is also what player i will judge the probability of $(\alpha_1, \ldots, \alpha_n)$ to be, if he is of type a_i. One way to justify this assumption is to suppose that the types of the players are determined by some well-understood stochastic process as, for example, if the players were selected by sampling at random out of populations with known proportions of each type.

(This consistency assumption is not really necessary for the structures which we are about to develop. In fact, we will henceforth refer only to the conditional probability functions P_i and the marginal probability functions R_i. A reader who objects to the consistency assumption may instead interpret these functions as follows: Let $P_i(\alpha_1, \ldots, \alpha_n|a_i)$ be the subjective probability which *player i* would assign to the types-vector $(\alpha_1, \ldots, \alpha_n)$ if he were type a_i, and let $R_i(a_i)$ be the prior marginal probability which the *arbitrator* would assign to the event that player i is type a_i. All of our results will still make sense when P_i and R_i are interpreted in this way.)

2. BAYESIAN INCENTIVE-COMPATIBILITY

In this section we restrict our attention to choice mechanisms using the standard response sets. That is, the arbitrator asks each player what his type is, and player i may respond by naming any of the possible types in A_i.

Since the arbitrator cannot force the players to give truthful responses, he must design the choice mechanism so that it does not give incentive for dishonesty. Given a choice mechanism π, for any player i and for any $a_i \in A_i$ and $b_i \in A_i$, let:

$$(5) \qquad Z_i(\pi, b_i|a_i) = \sum_{\boldsymbol{\alpha}\in A_i\times\ldots\times A_n} \sum_{c\in C} P_i(\boldsymbol{\alpha}|a_i)\pi(c|\boldsymbol{\alpha}_{-i}, b_i)U_i(c, \boldsymbol{\alpha})$$

where $(\boldsymbol{\alpha}_{-i}, b_i) = (\alpha_1, \ldots, \alpha_{i-1}, b_i, \alpha_{i+1}, \ldots, \alpha_n)$. (Recall that $P_i(\boldsymbol{\alpha}|a_i) = 0$ if $\alpha_i \neq a_i$.) Then $Z_i(\pi, b_i|a_i)$ is the conditionally-expected utility payoff for player i, given that his type is a_i, if he says that his type is b_i when π is the choice mechanism and when all other players are expected to tell the truth.

A choice mechanism π using the standard response sets is said to be *Bayesian incentive-compatible* if

$$(6) \qquad Z_i(\pi, a_i|a_i) \geqslant Z_i(\pi, b_i|a_i) \qquad \text{for all } i,\ a_i \in A_i,\ b_i \in A_i.$$

(D'Aspremont and Gerard-Varet [**1**] suggest the adjective "Bayesian" to distinguish this definition from the stronger definition of incentive-compatibility given by Hurwicz [**6**]. Here we may sometimes drop the adjective "Bayesian" and still mean (6), since no other definition of incentive-compatibility will be used in this paper.) If a choice mechanism is Bayesian incentive-compatible, then no player would expect any positive gains from being the only player to lie about his type when all others are planning to tell the truth. Thus, universal honesty is an equilibrium for the players if and only if the choice mechanism is Bayesian

incentive-compatible. (Recall that we are assuming that statements to the arbitrator are confidential; no player knows what the others are saying. So each player compares his conditionally-expected utilities, given only the information about his own true type, to find his best response.) So an arbitrator who selects a choice mechanism using the standard response sets cannot hope to get only honest responses unless he selects a Bayesian incentive-compatible mechanism. Otherwise, someone is bound to expect to profit from being the first to lie.

If choice mechanism π is used and if everyone is honest, then player i's conditionally-expected payoff when he knows a_i is

$$V_i(\pi|a_i) = Z_i(\pi, a_i|a_i). \tag{7}$$

The allocation of conditionally-expected payoffs associated with mechanism π is then the vector

$$\boldsymbol{V}(\pi) = ((V_1(\pi|a_1))_{a_1 \in A_1}, \ldots, (V_n(\pi|a_n))_{a_n \in A_n}). \tag{8}$$

(Thus $\boldsymbol{V}(\pi)$ is a vector or list of $\Sigma_{i=1}^{n} |A_i|$ real numbers, indexed on the disjoint union of the A_i sets.) The $\boldsymbol{V}(\pi)$ vector tells us what utility-expectation each player would enjoy in any of his possible types, if he learns that the arbitrator is planning to use choice mechanism π, and if all players are expected to respond honestly to the arbitrator.

If the arbitrator could use any choice mechanism and expect honest responses, then we would define the *feasible set* of expected allocation vectors to be

$$F = \{\boldsymbol{V}(\pi): \pi \text{ is a choice mechanism}\}. \tag{9}$$

Unfortunately, we know that honest responses cannot be expected unless the mechanism is Bayesian incentive-compatible. So we must restrict our attention to the subset of F which can be reached with Bayesian incentive-compatible mechanisms. The set of *incentive-feasible* expected allocation vectors is therefore defined to be

$$F^* = \{\boldsymbol{V}(\pi): \pi \text{ is Bayesian incentive-compatible}\}. \tag{10}$$

THEOREM 1: *F^* is a nonempty convex and compact subset of F.*

PROOF: A choice mechanism on the standard response sets is a real-valued function defined on $C \times A_1 \times \cdots \times A_n$ and satisfying

$$\sum_{c' \in C} \pi(c'|\boldsymbol{\alpha}) = 1 \quad \text{and} \quad \pi(c|\boldsymbol{\alpha}) \geq 0 \tag{2'}$$

for every $c \in C$ and every $\boldsymbol{\alpha} \in A_1 \times \ldots \times A_n$. Thus the set of choice mechanisms on the standard response sets is a compact convex subset of the vector-space of all real-valued functions on $C \times A_1 \times \ldots \times A_n$.

Notice also that, from (5), each $Z_i(\pi, b_i|a_i)$ depends linearly on π. Thus (6) is a collection of linear inequalities in π. So the set of Bayesian incentive-compatible mechanisms also forms a compact convex set, as it is the solution of the finite collection of linear inequalities in (2') and (6).

Of course, incentive-compatible mechanisms do exist. For example, letting $\pi(c|\boldsymbol{\alpha}) = 1/|C|$ for all c and $\boldsymbol{\alpha}$ would be incentive-compatible.

Now observe that $\boldsymbol{V}(\pi)$ is also a linear function of π. Thus F^* must be a nonempty compact convex set in allocation space, as it is the image, under a linear map, of the set of incentive-compatible mechanisms. *Q.E.D.*

It is well known that incentive-compatibility may be a significant restriction, in the sense that F^* may be a much smaller set that F. (See Hurwicz [**6**] and Rosenthal [**8**].) This fact has important implications for the arbitrator, as he tries to find a choice mechanism which will give the players high levels of expected utility. He certainly would not want to use a mechanism π if it were *strictly dominated* by another available mechanism π', in the sense that

$$V_i(\pi|a_i) < V_i(\pi'|a_i) \qquad \text{for all } i \text{ and all } a_i \in A_i. \tag{11}$$

But for many collective choice problems (see the example in Section 5), most of the undominated (weakly Pareto optimal) frontier of F may be outside of the incentive-feasible set F^*.

Since we cannot get allocations outside F^* with honest equilibria, we should now ask, could a choice mechanism have any equilibria (allowing some anticipated dishonesty) which would generate expected utility allocations outside F^*? In Section 3, we will argue that the answer to this question is no, that the players' equilibrium response behavior will always lead to an allocation in F^*, for any choice mechanism. Then, in Section 4 and 5, we will return to the problem of finding an expected utility allocation which is fair and efficient, subject to this "second best" restriction to F^*.

3. RESPONSE-PLAN EQUILIBRIA

In the preceding section we restricted our attention to choice mechanisms using the standard (A_i) response sets. We now turn to consider the general case. So throughout this section, we assume only that each player's response set S_i is a nonempty finite set.

A *response plan* for player i is a function σ_i mapping each type $a_i \in A_i$ onto a probability distribution over his response set S_i. That is, if σ_i is player i's response plan, then $\sigma_i(s_i|a_i)$ is the probability that player i will tell the arbitrator s_i if his true type is a_i. (Thus we must have $\sigma_i(s_i|a_i) \geq 0$ and $\Sigma_{s_i' \in S_i} \sigma_i(s_i'|a_i) = 1$ for all i and a_i.)

If $(\sigma_1, \ldots, \sigma_n)$ lists the players' response plans for the choice mechanism π, and if player i knows that a_i is his true type, then player i's expected utility payoff is

$$W_i(\pi, \sigma_1, \ldots, \sigma_n|a_i) = \sum_{\boldsymbol{\alpha} \in A_1 \times \ldots \times A_n} \sum_{s \in S_1 \times \ldots \times S_n} \sum_{c \in C} P_i(\boldsymbol{\alpha}|a_i) \cdot \left(\prod_{j=1}^{n} \sigma_j(s_j|a_j) \right) \cdot \pi(c|s) \cdot U_i(c, \boldsymbol{\alpha}). \tag{12}$$

The vector of conditionally-expected payoffs generated by $(\sigma_1, \ldots, \sigma_n)$ is

$$\boldsymbol{W}(\pi, \sigma_1, \ldots, \sigma_n) = (((W_i(\pi, \sigma_1, \ldots, \sigma_n | a_i))_{a_i \in A_i})_{i=1}^n). \tag{13}$$

(This is a vector with $\Sigma_{i=1}^n |A_i|$ components, indexed on the disjoint union of the A_i sets, like the $\boldsymbol{V}(\pi)$ vectors in Section 2.)

Notice that we cannot classify response plans as "honest" or "dishonest" in this section, since we have not defined any truthful response function for our abstract response sets S_i.

Following Harsanyi [**4**], we say that $(\sigma_1, \ldots, \sigma_n)$ is a *response-plan equilibrium* for the choice mechanism π if, for any player i and type $a_i \in A_i$, for every possible alternative response plan σ_i' for player i

$$W_i(\pi, \sigma_1, \ldots, \sigma_n | a_i) \geq W_i(\pi, \sigma_1, \ldots, \sigma_{i-1}, \sigma_i', \sigma_{i+1}, \ldots, \sigma_n | a_i). \tag{14}$$

That is, a collection of response plans forms an equilibrium if no player would ever expect to gain from unilaterally changing his plan.

We can now define the set of *equilibrium-feasible* expected allocation vectors to be

$$F^{**} = \{\boldsymbol{W}(\pi, \sigma_1, \ldots, \sigma_n): \pi \text{ is a choice mechanism, and } (\sigma_1, \ldots, \sigma_n) \text{ is a response-plan equilibrium for } \pi\}. \tag{15}$$

The arbitrator can deliver any expected allocation vector in F^{**} by committing himself to the corresponding choice mechanism π and by recommending to all the players that they use the σ_i response plans which generate this expected utility allocation. It is reasonable to expect the players to follow his recommendations since, by definitions of an equilibrium, no one can expect to do any better by using another response plan.

The central result of this section is that equilibrium-feasibility is *not* more general than incentive-feasibility defined in the preceding section. That is, for any response-plan equilibrium of any choice mechanism, there is an equivalent incentive-compatible mechanism giving all types of all players the same expected payoffs. Thus there will be no loss of generality in assuming that the arbitrator should select an incentive-compatible mechanism with the standard response sets.

THEOREM 2: $F^{**} = F^*$.

PROOF: This theorem could be proven as a corollary of Rosenthal's Theorem 3 in [**8**]. For completeness we present the entire proof here.

If $(\sigma_1, \ldots, \sigma_n)$ is a response-plan equilibrium for a mechanism π on $S_1, \ldots, S_n$, then we can define an equivalent choice mechanism π' on $A_1, \ldots, A_n$ by

$$\pi'(c|\boldsymbol{\alpha}) = \sum_{\boldsymbol{s} \in S_1 \times \ldots \times S_n} \pi(c|\boldsymbol{s}) \cdot \left(\prod_{i=1}^n \sigma_i(s_i|\alpha_i) \right).$$

It is easy to check that

$$\boldsymbol{V}(\pi') = \boldsymbol{W}(\pi, \sigma_1, \ldots, \sigma_n),$$

so that the allocations generated are the same. Furthermore, the equilibrium inequalities (14) for π imply the incentive-compatible inequalities (6) for π'. Thus $\boldsymbol{x} = \boldsymbol{W}(\pi, \sigma_1, \ldots, \sigma_n) \in F^{**}$ implies $\boldsymbol{x} = \boldsymbol{V}(\pi') \in F^*$. So $F^{**} \subseteq F^*$.

To verify the other inclusion, $F^* \subseteq F^{**}$, simply observe that, for any incentive-compatible mechanism π', the honest response plans $\sigma'_1, \ldots, \sigma'_n$ defined by

$$\sigma'_i(b_i|a_i) = \begin{cases} 1 & \text{if } b_i = a_i, \\ 0 & \text{if } b_i \neq a_i, \end{cases}$$

form a response-plan equilibrium for π'. So $\boldsymbol{x} = \boldsymbol{V}(\pi') \in F^*$ implies that $\boldsymbol{x} = \boldsymbol{W}(\pi', \sigma'_1, \ldots, \sigma'_n) \in F^{**}$. *Q.E.D.*

4. INCENTIVE-EFFICIENCY

Theorem 2 leads us back to the problem posed at the end of Section 2: What should the arbitrator do when the incentive-feasible set F^* is much smaller than the theoretically feasible set F (recall lines (9) and (10))? Because a dictatorship is always incentive-compatible (see Gibbard [**2**] and Satterthwaite [**9**]), some extreme points of the Pareto optimal frontier of F will always be incentive-feasible (that is, in F^*). Rosenthal [**8**] has shown that other regions of the Pareto optimal frontier of F may also be incentive-feasible for some cases, when special conditions are satisfied. But in general (as Example 3 in [**8**] and our example in Section 6 will show), for many Bayesian collective choice problems most of the Pareto optimal frontier of F may be outside of the incentive-feasible set F^*.

In particular, there is not much comfort for the arbitrator to know that making one player a dictator would be both incentive-compatible and Pareto optimal. After all, Pareto optimality is not the only criterion for judging choice mechanisms; the arbitrator also seeks to make an equitable compromise between the conflicting desires of the players.

Does this mean that the criterion of Pareto optimality cannot be applied to the Bayesian collective choice problem? We suggest that the answer to this question is no, that the Pareto optimality criterion is still relevant, but that it should be applied relative to F^* instead of F. It is unreasonable to base normative standards on comparisons with plans which are known to be unimplementable. It may be that some mechanism π' would give a high vector of expected payoffs $\boldsymbol{V}(\pi')$ if everyone were certain to be honest; but if π' is not incentive-compatible then someone is bound to find advantage from lying, and evaluations assuming universal honesty are just wishful thinking.

So we must restrict our attention to Bayesian incentive-compatible mechanisms. Any incentive-compatible mechanism which is dominated by another incentive-compatible mechanism ought to be ruled out, which leaves only the mechanisms which generate allocations on the undominated frontier of F^*. We shall refer to these mechanisms as *incentive-efficient.* That is, π is incentive-efficient if and only if it is a Bayesian incentive-compatible choice mechanism and is not strictly dominated (in the sense of (11)) by any other Bayesian incentive-compatible mechanism.

5. THE BARGAINING SOLUTION

Within the set of incentive-efficient mechanisms, the arbitrator still has a considerable range of mechanisms to consider. Some incentive-efficient mechanisms may be better for one player, and some may be better for another. So we must now ask: is there any natural or theoretically appealing way to select a unique incentive-efficient choice mechanism as the "solution" to our Bayesian collective choice problem? Harsanyi and Selten [**5**] derived a solution concept for a very similar class of problems, based on earlier work by Nash [**7**]. To apply their methods to our problem, however, we will need one further bit of structure: a *conflict outcome* c^* in C must be specified.

The conflict outcome represents what would happen by default if the arbitrator failed to lead the players to an agreement. In political choice problems, the conflict outcome could be a status quo which must prevail unless the players all agree otherwise. In market problems, the conflict outcome could be the no-trade position. In other applications, there may be a natural noncooperative game (a Bayesian game of the form described by Harsanyi [**4**]) which the players would have to play if they could not agree in the collective choice problem; in this case the conflict point should be some designated equilibrium of this noncooperative game.

Associated with the conflict outcome c^* is the *conflict payoff vector*,

$$\boldsymbol{t} = ((t_{a_1})_{a_1 \in A_1}, (t_{a_2})_{a_2 \in A_2}, \ldots, (t_{a_n})_{a_n \in A_n}), \tag{16}$$

where

$$t_{a_i} = \sum_{\boldsymbol{\alpha} \in A_1 \times \ldots \times A_n} P_i(\boldsymbol{\alpha} | a_i) \cdot U_i(c^*, \boldsymbol{\alpha}).$$

That is, each t_{a_i} number is player i's conditional expectation, given that a_i is his true type, of what his utility payoff would be if the conflict outcome occurred. Notice that a choice mechanism which always chooses c^* is trivially incentive-compatible. (No player would have any incentive to lie to an arbitrator whose plans called for c^* no matter how the players might respond.) So the conflict payoff vector generated by c^* is incentive-feasible; that is, $\boldsymbol{t} \in F^*$.

Given the conflict payoff vector $\boldsymbol{t}$ our collective choice problem becomes a *bargaining problem*, a generalized version of the bargaining problem originally studied by Nash [**7**], in that we have a feasible set F^* and a reference point $\boldsymbol{t}$ in F^*.

Let F^*_+ be the set of all incentive-feasible payoff vectors which are *individually rational*, in that no player of any type expects to do worse than in the conflict outcome:

$$F^*_+ = F^* \cap \{\boldsymbol{y} : y_{a_i} \geq t_{a_i} \text{ for all } i \text{ and all } a_i \in A_i\}. \tag{17}$$

Following Harsanyi and Selten [**5**], we define the *incentive-feasible bargaining*

solution to be the vector $x \in F^*_+$ which maximizes the *generalized Nash product*

$$(18) \qquad \prod_{i=1}^{n} \left(\prod_{a_i \in A_i} (x_{a_i} - t_{a_i})^{R_i(a_i)} \right)$$

over the set F^*_+.

THEOREM 3: *Suppose that c^* is not incentive-efficient (so that $\boldsymbol{t}$ is strictly dominated in F^*). Then there exists a unique incentive-feasible bargaining solution.*

PROOF: F^*_+ is compact, so the generalized Nash product does have a maximum point in F^*_+. As long as $\boldsymbol{t}$ is strictly dominated in F^*, we know that any maximum point $\boldsymbol{x}$ must strictly dominate $\boldsymbol{t}$, that is,

$$x_{a_i} > t_{a_i}, \quad \text{for every } i \text{ and } a_i \text{ in } A_i,$$

in order for the Nash product to be positive at $\boldsymbol{x}$. But then, the maximum point would also maximize the log of the Nash product

$$\sum_{i=1}^{n} \sum_{a_i \in A_i} R_i(a_i) \cdot \log (x_{a_i} - t_{a_i})$$

over the incentive-feasible allocations strictly dominating $\boldsymbol{t}$. The log of the Nash product is strictly convex in $\boldsymbol{x}$, so it can have at most one maximum point over the convex set of incentive-feasible allocations which strictly dominate $\boldsymbol{t}$. So the incentive-feasible bargaining solution must be unique. *Q.E.D.*

Harsanyi and Selten [**5**] derived this generalized Nash product criterion for the case of $n = 2$. Their theory rests on a series of eight axioms which they suggest a bargaining solution should satisfy, including: individual rationality, symmetry over players and types, efficiency, and invariance under several inessential ways of transforming a bargaining problem. The main distinction between our incentive-feasible bargaining solution and their bargaining solution (besides our straightforward generalization to general n-person problems) lies in the way the feasible sets are defined. In Part II of [**5**], Harsanyi and Selten define their feasible set as the convex hull of the payoff allocations generated by some special equilibria (the *strict* equilibria) of a particular choice mechanism which they describe. By Theorem 2, their feasible set is a subset of our incentive-feasible set F^*, which may be much larger.

Any incentive-compatible mechanism π which generates the incentive-feasible solution (in the sense that $\boldsymbol{x} = \boldsymbol{V}(\pi)$) is called an *implementation* of the solution. Of course, the solution must have an implementation, since the solution is in F^*, which by definition is the incentive-feasible set. There may be several implementations of the solution, but since they all generate the same expected payoff allocation, they are all essentially equivalent in the eyes of all players. The solution must be undominated in F^* (since it maximizes (18), which is an increasing function of $\boldsymbol{x}$), so an implementation of the incentive-feasible solution will always be incentive-efficient.

6. EXAMPLE

To illustrate these ideas, consider a simple collective choice problem, involving two players who must share the cost of a public works project which would benefit them both. The project (perhaps a new pavement for a road which only these two players use) would cost \$100, and the two players have called in an arbitrator to help them divide the cost. The arbitrator knows that the project would be worth \$90 to player 2, but its value to player 1 would depend on his type. If player 1 is of type 1.0, then the project is also worth \$90 to him, but if player 1 is of type 1.1 ("thinks old unpaved roads have rustic charm") then the project is worth only \$30 to him. Only player 1 knows for sure what his type is, but the arbitrator and player 2 figure that type 1.0 is much more likely, so $R_1(1.0) = .9$ and $R_1(1.1) = .1$.

Thus, no matter what player 1's type is, the project appears to be worth more than it costs: either it is worth $90 + 90 = 180$ if player 1 is of type 1.0, or it is worth $30 + 90 = 120$ if player 1 is of type 1.1. If the decision to produce the project could be made separately from the decision on allocating the cost, then it would seem clear that the project should be undertaken. Unfortunately, these decisions cannot be separated without either violating incentive-compatibility or being very unfair to player 2. Thus, we shall see that the incentive-feasible bargaining solution will give a small but positive probability to *not* undertaking the project.

Before we undertake a formal analysis of this problem, it may be worthwhile to informally survey a few choice mechanisms which unaided intuition might suggest. For example, one might suggest that the players should pay equally (\$50 each) for the project if 1 is type 1.0 (so that each enjoys the same net gain of $90 - 50 = \$40$), but that 1 should pay \$20 and 2 should pay \$80 if 1 is type 1.1 (so that each enjoys the same net gain of \$10). Unfortunately, this choice mechanism is not incentive-compatible, since 1 could do better by saying he was type 1.1 even if 1.0 were true.

To guarantee incentive-compatibility, one might also suggest a mechanism which does not use any response information from player 1. For example one might suggest that 1 should pay \$47 and 2 should pay \$53 to finance the project, regardless of 1's type. This mechanism is incentive-compatible, but it is not individually rational (assuming the conflict outcome is "do not undertake the project"), since if player 1 is type 1.1 then he will be made worse off in paying \$47 for a project which is worth only \$30 to him. It is no comfort to player 1 to know that this would be a good deal if he were type 1.0 when he knows that he is type 1.1. A mechanism which uses no response information from player 1 would be individually rational for player 1 only if player 1's "flat rate" fee is \$30 or less; but this would require player 2 to pay \$70 or more, which would seem very unfair if 1 is type 1.0 (as is 90 per cent likely).

One choice mechanism which is both individually rational and incentive-compatible is as follows: Ask player 1 what his type is; if he says 1.0 then charge each player \$50 to produce the project; if he says 1.1 then do not produce the project. If player 1 is type 1.0 then his expected net gain is \$40 from this plan; if player 1 is type 1.1 then his net gain is \$0, and player 2's expected net gain is $.9(\$40) + .1(\$0) = \$36$. This mechanism is not incentive-efficient however,

because there are other incentive-compatible mechanisms making all types better off. To find these incentive-efficient mechanisms and compute the incentive-feasible bargaining solution, we must first translate our problem into a formal collective choice problem of the form (1).

To formally model this collective choice problem, let $C = \{c_0, c_1, c_2\}$, $A_1 = \{1.0, 1.1\}$, and $A_2 = \{2\}$. We have $P(1.0, 2) = .9$ and $P(1.1, 2) = .1$, and the utility functions are as follows:

(u_1, u_2)	c_0:	c_1:	c_2:
$a_1 = 1.0$:	$(0, 0)$	$(-10, 90)$	$(90, -10)$
$a_1 = 1.1$:	$(0, 0)$	$(-70, 90)$	$(30, -10)$

The choices are interpreted as follows: c_0 is the choice "do not undertake the project"; c_1 is the choice "undertake the project and make player 1 pay for it"; and c_2 is the choice "undertake the project and make player 2 pay for it". There is no need to include choices in C to represent the intermediate financing options between c_1 and c_2, because they can be represented by the "randomized" strategies. For example, to undertake the project by charging player 1 \$40 and player 2 \$60 would give the two players the same expected utility as the randomized strategy $.4c_1 + .6c_2$ (giving c_1 probability $\pi(c_1) = .4$ and c_2 probability $\pi(c_2) = .6$). (In this simple example we are assuming that the players have utility which is linear in money.)

The natural conflict outcome for this problem is $c^* = c_0$; that is, the project will not be produced if the players cannot reach an agreement.

To keep our formulas concise as we describe choice mechanisms, we shall use the abbreviations

$$\pi_j^0 = \pi(c_j|1.0, 2) \qquad \text{and} \qquad \pi_j^1 = \pi(c_j|1.1, 2)$$

for a randomized choice mechanism π.

With this notation, the incentive-compatible choice mechanisms are those satisfying the following inequalities:

$$\begin{aligned} &-10\pi_1^0 + 90\pi_2^0 \geq -10\pi_1^1 + 90\pi_2^1, \\ &-70\pi_1^1 + 30\pi_2^1 \geq -70\pi_1^0 + 30\pi_2^0, \\ &\pi_0^0 + \pi_1^0 + \pi_2^0 = 1, \quad \pi_0^1 + \pi_1^1 + \pi_2^1 = 1, \end{aligned} \tag{19}$$

and all $\pi_j^i \geq 0$. The first inequality says that player 1 should not want to claim to be type 1.1 if he is really type 1.0; the second inequality says that 1 should not want to claim to be 1.0 if he is really 1.1. The other conditions in (19) merely state that the choice mechanism π selects a proper probability distribution over C for each possible announcement.

Then the incentive feasible set F^* is the set of allocation vectors $\boldsymbol{x} = (x_{1.0}, x_{1.1}, x_2)$ such that

$$
\begin{aligned}
& x_{1.0} = 0\pi_0^0 - 10\pi_1^0 + 90\pi_2^0, \\
& x_{1.1} = 0\pi_0^1 - 70\pi_1^1 + 30\pi_2^1, \qquad \text{and} \\
& x_2 = .9(0\pi_0^0 + 90\pi_1^0 - 10\pi_2^0) + .1(0\pi_0^1 + 90\pi_1^1 - 10\pi_2^1),
\end{aligned} \tag{20}
$$

where π satisfies (19).

Analysis of (19) and (20) can show that F^* is just the convex null of the following five allocation vectors:

$(0, 0, 0)$ (implemented by $\pi_0^0 = \pi_0^1 = 1$);

$(-10, -70, 90)$ (implemented by $\pi_1^0 = \pi_1^1 = 1$);

$(90, 30, -10)$ (implemented by $\pi_2^0 = \pi_2^1 = 1$);

$(0, 0, 72)$ (implemented by $\pi_1^0 = .9$, $\pi_2^0 = .1$, $\pi_0^1 = 1$);

$(60, 0, 18)$ (implemented by $\pi_1^0 = .3$, $\pi_2^0 = .7$, $\pi_0^1 = 1$).

Of these five vectors, the first and last are dominated by combinations of the other three, so the undominated frontier of F^* is the convex hull of the middle three.

The incentive-feasible bargaining solution is the solution to the non-linear programming problem:

maximize $((x_{1.0})^{.9} \cdot (x_{1.1})^{.1} \cdot x_2)$

subject to x and π satisfying (19) and (20).

Using the Kuhn–Tucker conditions, it can be shown that the incentive-feasible bargaining solution is

$$x_{1.0} = 39.5, \qquad x_{1.1} = 13.2, \qquad x_2 = 36.$$

This solution is implemented by

$$\pi_1^0 = .505, \qquad \pi_2^0 = .495, \qquad \pi_0^1 = .561, \qquad \text{and} \qquad \pi_2^1 = .439.$$

That is, if player 1 claims to be type 1.0, then the project is produced for sure and its cost is split approximately equally between the two players (with 1 paying slightly more than 2). If 1 claims to be type 1.1 then the project is produced only with probability .439, but if it is produced then player 2 pays the entire cost.

Our incentive-feasible bargaining solution is incentive-efficient, even though there is a 5.61 per cent chance ($\pi_0^1 \times R_1(1.1) = .0561$) that the project might not be undertaken. The project would always be worthwhile *if* its costs could be divided fairly and incentive-compatibly. But, as we showed informally in our earlier

discussion of selected mechanisms, there is no incentive-compatible way to guarantee the project without either hurting player 1 if he is type 1.1, or else being unfair to player 2.

Northwestern University

Manuscript received June, 1977; revision received February, 1978.

REFERENCES

[1] D'ASPREMONT, C., AND L. GERARD-VARET: "Incentives and Incomplete Information," CORE Discussion Paper No. 7705, March, 1977.
[2] GIBBARD, A.: "Manipulation of Voting Schemes: A General Result," *Econometrica*, 41 (1973), 587–602.
[3] ———: "Manipulation of Schemes That Mix Voting With Chance," *Econometrica*, 45 (1977), 665–681.
[4] HARSANYI, J. C.: "Games with Incomplete Information Player by 'Bayesian' Players," *Management Science*, 14 (1967–8), 159–189, 320–334, 486–502.
[5] HARSANYI, J. C., AND R. SELTEN: "A Generalized Nash Solution for Two-Person Bargaining Games with Incomplete Information," *Management Science*, 18 (1972), P80–P106.
[6] HURWICZ, L.: "On Informationally Decentralized Systems," in *Decision and Organization*, ed. by R. Radner and B. McGuire. Amsterdam: North-Holland Press, 1972, pp. 297–336.
[7] NASH, J. F.: "The Bargaining Problem," *Econometrica*, 18 (1950), 155–162.
[8] ROSENTHAL, R. W.: "Arbitration of Two-Party Disputes Under Uncertainty," forthcoming in *Review of Economic Studies*.
[9] SATTERTHWAITE, M. A.: "Strategy-Proofness and Arrow's Conditions: Existence and Correspondence Theorems for Voting Procedures and Social Welfare Functions," *Journal of Economic Theory*, 10 (1975), 187–217.