R

Discussion Paper No. 481

MECHANISM DESIGN BY AN INFORMED PRINCIPAL

by

Roger B. Myerson

June 1981
revised, September 1982

J. L. Kellogg Graduate School of Management
Northwestern University
Evanston, Illinois 60201

Abstract. When a principal with private information designs a mechanism to coordinate his subordinates, he faces a dilemma: to conceal his information, his selection of mechanism must not depend on his information; but his information may influence which mechanism he prefers. To resolve this dilemma, this paper develops a theory of inscrutable mechanism selection. The principal's neutral optima are defined as the smallest possible set of unblocked mechanisms. They are shown to exist and are characterized using parametric linear programs. Any safe and undominated mechanism is a neutral optimum. Any neutral optimum is an expectational equilibrium and a core mechanism.

Acknowledgements. Research for this paper was supported by the Kellogg Center for Advanced Study in Managerial Economics and Decision Sciences, and by a research fellowship from I.B.M. Many helpful suggestions and criticisms by colleagues in the M.E.D.S. department and the Math Center at Northwestern University and in the I.M.S.S.S. summer workshop at Stanford University are gratefully acknowledged.

1. Introduction

An individual who bargains when he has private information often faces a dilemma. On the one hand, he may want to conceal his information from the people with whom he is bargaining. But on the other hand, his goals in bargaining may depend on his information. For example, if the seller of a used car knows that he has a low quality car (a "lemon"), then he wants to conceal this fact from the buyer. But the seller's information may also make him prefer not to offer any warranty on the car's performance, even if he has to concede a lower price to do so. Should such a seller try to avoid giving any warranty, as he bargains with the buyer over the terms of sale, or should he offer a warranty, to conceal his information? The seller's actual preferences are in conflict with his need to be inscrutable. In this paper we shall develop a general theory of how individuals may resolve such conflicts.

Even in games with complete information, there is still no general definitive theory of bargaining between two individuals. We might expect the players to agree on some outcome on their Pareto frontier, but which point is agreed upon may depend on many factors. (See Roth [1979] for a survey of mathematical theories of bargaining.) However, if we assume that one player has all of the bargaining ability, then the solution to the bargaining problem with complete information is obvious: the individual with all the bargaining ability should insist on the best outcome possible for himself, subject to the constraint that the other individual cannot be made worse off than if he refused to cooperate.

Because the issues of bargaining with incomplete information are so complicated, a good research strategy is to begin by just studying this case, where one individual has all of the bargaining ability. Even in this case, which is trivial with complete information, difficult issues arise when the individual in control has private information. This paper will be devoted entirely to the study of this case. However, the insights which we develop here will also lay the foundations for later papers that will develop a general theory of bargaining with incomplete information between individuals who all have bargaining ability. (See Myerson [1982b] and [1982c].)

We shall refer to the individual with all of the bargaining ability as the <u>principal</u> in the bargaining situation, and all other individuals are the <u>subordinates</u>. This terminology is meant to suggest one kind of environment (the hierarchical organization) in which individuals typically do interact with asymmetric bargaining ability. Also, one salient feature of most principal-agent models (see, for example Ross [1973], Mirrlees [1976], Harris and Raviv [1979], and Holmström [1979]) is that the principal can implement the coordination mechanism that is best for him, subject to the constraint that the agent must be given at least the minimal incentives to act as the principal desires. It is this feature of the principal's role which we are generalizing in this paper. The idea of giving one individual the authority to select an incentive-compatible mechanism in general Bayesian social choice problems was suggested by Harris and Townsend [1981].

If an individual is the <u>principal</u>, in the sense of this paper, it does not mean that he can force the subordinates to do anything he wants. The subordinates may have control over private decisions which the principal cannot dictate, or they may have private information which the principal cannot observe, or they may simply have the option of leaving the principal's organization if he does not offer them some minimal expected payoffs. If an individual is the principal, it means that he has effective control over the channels of communication between all individuals and that the subordinates cannot make threats against him. That is, the principal knows that the subordinates will do whatever he asks, provided that he makes it at least minimally in their best interests to do so, so they cannot bargain against him for a larger share of the social surplus. If the principal designs a game for the subordinates to play, then he can be confident that they will use the strategies which he suggests for them in this game, provided that these strategies form a Nash equilibrium. Or, using his control over the channels of communication, the principal can direct the subordinates to use some correlated equilibrium, in the sense of Aumann [1974]. In general, when we say that an individual is the principal, we mean that he can control the subordinates only to the extent that he can manipulate their incentives, but they accept such manipulation passively.

One source of informed-principal problems is in the theory of signalling in markets with adverse selection. For example, Rothschild and Stiglitz [1976] and Wilson [1977] have studied equilibria in insurance markets where each customer has private information about his risk category, information which may affect the expected profits of an insurance company that sells him a policy. In a market equilibrium, Rothschild and Stiglitz and Wilson assume that competition between insurance companies should always give them zero expected profits. We may now ask, if there were just one customer bargaining with one insurance company, as monopsonist and monopolist, would they negotiate the same insurance policy as in a Rothschild-Stiglitz or a Wilson market equilibrium? Since the insurance company makes zero expected profits in the market model, clearly the bargaining model can simulate the market model only if the customer has all the bargaining ability; that is, the customer must be the principal bargainer. Examples are known in which the principal's neutral mechanisms, as defined in this paper, do coincide with Wilson's E2 equilibria (or <u>anticipatory equilibria</u>, in the sense of Riley

[1979]), when the customer is the principal. It is hoped that this equivalence might be shown to hold for some class of signalling problems. (This question has been investigated recently by Bhattacharya [1981].) Other related models of markets in which informed individuals are given price-setting power have been studied by Wilson [1980].

The basic structure of our model is developed in Section 2 of this paper. In Section 3, the principal's mechanism-selection problem is introduced. We argue that all types of the principal should be expected to select the same mechanism, even though they have different preferences, so that the selection itself does not reveal any information. In Section 4 we argue that, if there exists a mechanism that is safe and undominated (as will be defined), then it is essentially unique for this property, and all types of the principal should implement it. We call such safe and undominated mechanisms strong solutions. Sections 5 and 6 introduce the concepts of expectational equilibria and core mechanisms, to help delimit the set of mechanisms that the principal could reasonably consider, in cases where no strong solution exists.

In Section 7, we systematically approach the problem of developing a theory to determine what an informed principal should do. We define the principal's neutral optima as the set of mechanisms that cannot be blocked, with any concept of blocking that satisfies four axioms. The main results of this paper are the characterization of neutral optima, presented in Section 8, and the general proof of existence of neutral optima, from which the existence of expectational equilibria and core mechanisms is also derived.

Most of the technical proofs are deferred to Section 9.

2. Bayesian Incentive Problems and Incentive-Compatible Mechanisms

We consider a general Bayesian incentive problem with n individuals, numbered $i = 1,2,\ldots,n$. As in Myerson [1982a], we allow for both informational (adverse selection) and strategic (moral hazard) constraints on the ability of these individuals to coordinate themselves, so that our model can subsume the most general class of problems.

For each individual i, we let T_i denote the set of possible types for individual i. Each type t_i in T_i completely specifies some possible state of i's preferences, abilities, and beliefs. That is, i's type is a random variable which subsumes all of i's information that is not public knowledge. (This terminology is based on the seminal paper of Harsanyi [1967-8].) We shall assume that, from the first point in time when these n individuals can actually make decisions or interact with each other, each individual already knows his own type.

A mechanism is any rule determining the individuals' actions as a function of their types. The set of feasible mechanisms is limited by two factors. First, each individual must be given the incentive to report his private information honestly. That is, we assume that the individuals' types are unverifiable, so that each individual may conceal or lie about his type unless he is given the correct incentives to tell the truth. Second, each individual may control some private decisions that cannot be cooperatively coordinated with the others. These private decisions may be unverifiable, like the agent's level of effort in the conventional principal-agent problem; or these private decisions may be intrinsically unalienable, like a worker's option to refuse employment if compensation is below his reservation wage. In either case, the result is that there are some decisions or actions which cannot be implemented unless the individual responsible is given the correct incentives to choose them.

Thus, we must distinguish between actions that are publicly observable and enforceable, and actions that must be privately controlled. We let D_0 denote the set of all possible enforceable or public actions, which can be contractually specified. That is, any d_0 in D_0 represents a combination of actions and decisions which the individuals can (in principle) commit themselves to carry out, even if it may turn out ex post to be harmful to any or all of the individuals. For each individual i, we let D_i represent the set of all possible private actions controlled by individual i. For example, D_i may be a set of unobservable effort levels for individual i, and D_0 may be a set of capital-resource allocations for the n individuals.

We let

$$T = T_1 \times \ldots \times T_n$$

denote the set of all possible combinations of individuals' types, with $t = (t_1,\ldots,t_n)$ denoting a typical types-vector or state in T. We let T_{-i} denote the set of possible combinations of types of the individuals other than i, that is

$$T_{-i} = T_1 \times \ldots \times T_{i-1} \times T_{i+1} \times \ldots \times T_n.$$

Similarly, we let

$$D = D_0 \times D_1 \times \ldots \times D_n$$

denote the set of all possible combinations of public and private actions, with $d = (d_0,d_1,\ldots,d_n)$ denoting a typical actions-vector or outcome in D. For mathematical convenience, we shall assume that D and T are (nonempty) finite sets.

Given any vector of types t and actions d, we let $u_i(d,t)$ denote the payoff to individual i, measured in a vonNeumann-Morgenstern utility scale, when d is the outcome and t is the actual state of the game. We let $p_i(t_{-i}|t_i)$ denote the conditional probability that individual i would assign to the event that $t = (t_1,\ldots,t_n)$ is the actual state of the game, given that he knows his actual type to be t_i. (We use here the notation $t_{-i} = (t_1,\ldots,t_{i-1},t_{i+1},\ldots,t_n)$.) As a regularity assumption, we will assume that these conditional probabilities are all nonzero, so that

(2.1) $$p_i(t_{-i}|t_i) > 0, \quad \forall i\varepsilon\{1,\ldots,n\}, \quad \forall t\varepsilon T.$$

That is, no individual is absolutely sure that any combination of the others' types is impossible. (This assumption will be needed in definition (5.1) and in the proof of Lemma 2.)

Thus, the general Bayesian incentive problem Γ is characterized by these structures

(2.2) $$\Gamma = (D_0,D_1,\ldots,D_n,T_1,\ldots,T_n,u_1,\ldots,u_n,p_1,\ldots,p_n).$$

Our next task is to describe the set of feasible mechanisms for coordinating the public and private actions, as a function of the individuals' types.

Consider the following scenario. Each individual simultaneously and confidentially reports his type to a trustworthy mediator (or a mechanical information processor). The mediator then chooses an outcome $d = (d_0,d_1,\ldots,d_n)$ in D, as a (possibly random) function of the vector of types reported to him. Then the enforceable action d_0 is carried out, and each individual is confidentially informed that d_i is the private action recommended for him.

Formally, a <u>mechanism</u> is any function $\mu:D \times T \to \mathbb{R}$ such that

(2.3) $$\sum_{c\varepsilon D} \mu(c|t) = 1 \text{ and } \mu(d|t) \geq 0, \quad \forall d\varepsilon D, \forall t\varepsilon T.$$

Here $\mu(d|t)$ is interpreted as the probability that d will be the outcome chosen by the mediator in the above scenario, if t is the reported state of individuals' types.

For any possible types t_i and s_i of individual i, any function $\delta_i : D_i \to D_i$, and any mechanism μ, we make the following definitions:

$$(2.4) \qquad U_i(\mu|t_i) = \sum_{t_{-i}\varepsilon T_{-i}} p_i(t_{-i}|t_i) \sum_{d\varepsilon D} \mu(d|t)\, u_i(d,t),$$

and

$$(2.5) \qquad U_i^*(\mu,\delta_i,s_i|t_i) =$$

$$= \sum_{t_{-i}\varepsilon T_{-i}} p_i(t_{-i}|t_i) \cdot \sum_{d\varepsilon D} \mu(d|t_{-i},s_i)\, u_i((d_{-i},\delta_i(d_i)),t).$$

(In this paper, whenever t, t_i, and t_{-i} appear in the same formula, t_{-i} denotes the vector of all components other than t_i in the vector t = $(t_1,\ldots,t_n)$. Also, (t_{-i},s_i) and $(d_{-i},\delta_i(d_i))$ are respectively the vectors that differ from t and d in that s_i replaces t_i and $\delta_i(d_i)$ replaces d_i.) Thus, $U_i(\mu|t_i)$ is the conditionally expected utility for individual i, given that his type is t_i, if all individuals report their types honestly and carry out their recommended private actions obediently, when the mediator uses mechanism μ. On the other hand, if individual i reports s_i and plans to use private action $\delta_i(d_i)$ when d_i is recommended, while all other individuals are honest and obedient, then $U_i^*(\mu,\delta_i,s_i|t_i)$ is i's conditionally expected utility from mechanism μ, given that i's true type is t_i. Notice that the mediator's recommendation may convey information to i about the others' types, so that i might rationally choose his private actions as some function $\delta_i(\cdot)$ of his recommended action.

The mechanism μ is <u>incentive compatible</u> (in the Bayesian sense of D'Aspremont and Gerard-Varet [1979]) iff

$$(2.6) \qquad U_i(\mu|t_i) \geq U_i^*(\mu,\delta_i,s_i|t_i), \quad \forall i\varepsilon\{1,\ldots,n\},\ \forall t_i\varepsilon T_i,\ \forall s_i\varepsilon T_i,\ \forall\delta_i : D_i \to D_i.$$

Condition (2.6) aserts that honest and obedient participation in the mechanism μ must be a Bayesian Nash equilibrium for the n individuals, in the sense of Harsanyi [1967-8]. In Myerson [1982a], it has been shown that there is no loss of generality in considering only incentive-compatible coordination mechanisms, in the following sense: for any Bayesian equilibrium of any other coordination game which the individuals might play, there exists an equivalent incentive-compatible mechanism satisfying (2.6). This idea, called the <u>revelation principle</u>, has been presented in related contexts by Gibbard [1973], Rosenthal [1978], Dasgupta, Hammond, and Maskin [1979], Holmstrom [1977], Harris and Townsend [1981], and Myerson [1979].

One special case of the above structures may be worth considering, as an example. Suppose that each individual's set of private actions is simply $D_i = \{\text{"accept"}, \text{"reject"}\}$, and that all utility payoffs will be zero if any individual chooses his "reject" option. Suppose also that there is an enforceable action ("fire everyone") that also makes all payoffs zero. Then without loss of generality, we need only consider mechanisms in which no individual is ever asked to "reject", since the "fire everyone" action may be used instead. Then the incentive constraints (2.6) reduce to

$$(2.7) \qquad U_i(\mu|t_i) \geq \sum_{t_{-i}\in T_{-i}} \sum_{d\in D} p_i(t_{-i}|t_i)\,\mu(d|t_{-i},s_i)\,u_i(d,t), \quad \forall i,\ \forall t_i\in T_i,\ \forall s_i\in T_i,$$

and

$$(2.8) \qquad U_i(\mu|t_i) \geq 0, \quad \forall i,\ \forall t_i\in T_i.$$

That is, no individual should have any incentive to lie or reject in the mechanism.

3. The Inscrutable Principal

If an outsider with no private information (an academic economist, perhaps) were given the authority to control all communication between the n individuals and to determine the enforceable actions in D_0, then he could implement any incentive-compatible mechanism satisfying constraints (2.3) and (2.6). But if one of the n informed individuals can influence the selection of the mechanism when he already knows his own type, then a fundamentally new issue arises to constrain the choice of mechanism: if the selection of the coordination mechanism depends in any way on one individual's type, then the selection of the mechanism itself will convey information about his type to the other individuals. Under these circumstances, for a mechanism to be feasible, it must be incentive compatible after all other individuals have inferred whatever information might be implicit in the establishment of the mechanism itself.

In this paper we will assume that individual #1 can effectively control all communications and can dictate how the action in D_0 is to be determined, without any need to bargain or compromise with any of the other n-1 individuals. (The difference between D_0 and D_1 is that the action in D_1 is subject to moral hazard, in the incentive constraints, but the action in D_0 is not.) That is, individual 1 has complete authority to select any mechanism for coordinating the enforceable and private actions of the n individuals. The mediator described in the preceding section is a mere tool of individual 1, implementing the coordination mechanism that he selects. In view of this asymmetry of power, we shall henceforth refer to individual 1 as the principal in the system; individuals 2,...,n will be referred to as the subordinates.

We assume that the principal already knows his type at the time when he selects the mechanism, and that this is not a repeated situation. Thus, the best incentive-compatible mechanism for the principal maximizes his conditionally expected utility $U_1(\mu|t_1)$ given his true type t_1, subject to the constraints (2.3) and (2.6). But if the principal chooses μ to maximize $U_1(\mu|t_1)$, then his choice will depend on the true type t_1, and so the subordinate individuals may be able to infer something about the principal's type from his choice of μ. With this new information, the subordinates may find new opportunities to gain by dishonesty or disobedience. So a mechanism might not be incentive-compatible in practice, even though it satisfies (2.5), if the fact that μ is used allows the subordinates to learn about the principal's type.

Let R be any nonempty subset of T_1. We say that a mechanism μ is <u>incentive compatible given R</u> iff μ is incentive compatible for the principal (that is, μ satisfies (2.6) for i=1) and

$$(3.1) \qquad \sum_{\substack{t_{-i}\varepsilon T_{-i} \\ t_1 \varepsilon R}} \sum_{d\varepsilon D} p_i(t_{-i}|t_i)\,\mu(d|t)\,u_i(d,t)$$

$$\geqslant \sum_{\substack{t_{-i}\varepsilon T_{-i} \\ t_1 \varepsilon R}} \sum_{d\varepsilon D} p_i(t_{-i}|t_i)\,\mu(d|t_{-i},s_i)\,u_i((d_{-i},\delta_i(d_i)),\,t)$$

$$\forall i\varepsilon\{2,\ldots,n\},\ \forall t_i \varepsilon T_i,\ \forall s_i \varepsilon T_i,\ \forall \delta_i : D_i \rightarrow D_i.$$

(The summations in (3.1) indicate that t_{-i} is to range over vectors such that the first component t_1 is in R.) This condition (3.1) asserts that no subordinate i should expect to gain by reporting s_i and by disobeying his instructions according to δ_i, when he knows that t_i is his true type and that the principal's type is in R. Thus, if the subordinates expected that the principal would propose mechanism μ if his type were in R, but otherwise would propose some other mechanism, then μ could be successfully implemented only if it were incentive compatible given R. (The two sides of (3.1) differ from conditionally-expected utilities because we have not divided by i's probability of the event $t_1 \varepsilon R$; however, this factor can be ignored, since it is the same on both sides, and is positive by (2.1).)

This concept of conditional incentive compatibility describes what the principal could achieve if some information were revealed. However, as we try to construct a theory to determine which mechanism the principal should implement, there is no loss of generality in assuming that all types of the principal should choose the same mechanism, so that his actual choice of mechanism will convey no information. We may refer to this claim as the principle of inscrutability. Its essential justification is that the principal should never need to communicate any information to the subordinates by his choice of mechanism, because he can always build such communication into the process of the mechanism itself (in that $\mu(d|t)$ can depend on t_1).

A more formal justification for this principle of inscrutability may be given as follows. Suppose to the contrary, that there are some mechanisms $\{\mu_1,\ldots,\mu_K\}$ and sets of types $\{R_1,\ldots,R_K\}$ forming a partition of T_1, such that the types in R_k are expected to implement μ_k, for every k in $\{1,\ldots,K\}$. (For simplicity, we ignore randomized mechanism-selection plans here, but our argument could be easily extended to cover this case as well.) Since the subordinates would rationally infer that the principal's type is in R_k when μ_k is proposed, each μ_k must be incentive compatible given R_k. Since the principal already knows his type, he would choose to implement these mechanisms in this fashion only if they satisfy

$$(3.2) \qquad U_1(\mu_k|t_1) \geq U_1(\mu_j|t_1), \quad \forall j,\ \forall k,\ \forall t_1 \varepsilon R_k,$$

and are incentive compatible for him separately. But now consider the mechanism μ^* defined by

(3.3) $\quad \mu^*(d|t) = \mu_k(d|t)$ if $t_1 \varepsilon R_k$.

This mechanism μ^* is completely equivalent to the system of mechanisms $\{\mu_1,\ldots,\mu_K\}$ on the partition $\{R_1,\ldots,R_K\}$, giving the same distribution of outcomes in every state. That is, saying that "for each k, if the principal's type is in R_k then he will implement μ_k" is empirically indistinguishable from saying that "the principal will implement μ^*, no matter what his type is." It is straightforward to verify that μ^* is incentive compatible, using (3.1) (with $\mu = \mu_k$ and $R = R_k$) and (3.2) to prove that (2.6) holds for $\mu = \mu^*$.

The goal of this paper is to develop a theory to predict which mechanisms a principal with private information might select. For inscrutability, any mechanism that we predict must be reasonable for all of his types to select. If the principal's different types would actually prefer different incentive-compatible mechanisms, then the predicted mechanism must be some kind of compromise between the different goals of the principal's possible types. The main task of this paper is to develop formal notions of what such a "reasonable compromise" should be.

4. Safe and Undominated Mechanisms

We say that a mechanism μ is <u>dominated</u> by another mechanism ν iff $U_1(\mu|t_1) \leq U_1(\nu|t_1)$ for every t_1 in T_1, with strict inequality for at least one t_1 in T_1. We say that μ is <u>undominated</u> iff μ is incentive compatible and μ is not dominated by any other incentive-compatible mechanism. If $U_1(\mu|t_1) < U_1(\nu|t_1)$ for every t_1 in T_1, then μ is <u>strictly</u> dominated by ν.

Because the principal has effective control over the communication channels between himself and the subordinates, he should never be expected to implement any strictly dominated mechanism. To see why, suppose to the contrary that the principal is expected to implement some mechanism μ that is strictly dominated by another incentive-compatible mechanism ν. Then the principal could address the subordinates as follows:

> "I am going to implement the mechanism ν. Notice that all of my types prefer ν over μ, which you might have thought we would implement. Thus, you should not infer anything about my type from the fact that I have chosen ν rather than μ. With no new information about my type, you should each find it optimal to participate honestly and obediently in this incentive-compatible mechanism ν."

When we assume that the principal can communicate effectively and has all of the bargaining ability, we mean that the subordinates would understand such an argument and accept it.

We say that a mechanism μ is <u>safe</u> iff, for every type t_1 in T_1, μ is incentive compatible given $\{t_1\}$. That is, a safe mechanism is one which would be incentive compatible if the subordinates knew the principal's type. No matter what the subordinates might infer about the principal's type, he can successfully implement a safe mechanism, because it is incentive compatible given any subset of T_1.

Safe mechanisms may not necessarily exist for a Bayesian incentive problem. Even if one does exist, it may be strictly dominated in the class of incentive-compatible mechanisms. However, we now show that a mechanism that is both safe and undominated, if it exists, should be implemented by all types of the principal. This result defines a class of problems in which it is clear what the informed principal should do. We call a safe and undominated mechanism a strong solution for the principal.

Theorem 1. Suppose that μ is a strong solution. Let ν be any other mechanism, and let

$$S = \{t_1 \varepsilon T_1 \mid U_1(\nu|t_1) > U_1(\mu|t_1)\}.$$

If $S \neq \emptyset$ then ν is not incentive compatible given S. Furthermore, if $\hat{\mu}$ is any other safe and undominated mechanism, then

$$U_1(\hat{\mu}|t_1) = U_1(\mu|t_1), \quad \forall t_1 \varepsilon T_1.$$

Proof Consider the mechanism μ^* defined by

$$\mu^*(d|t) = \begin{cases} \nu(d|t) \text{ if } t_1 \varepsilon S, \\ \mu(d|t) \text{ if } t_1 \not\varepsilon S. \end{cases}$$

If $S \neq \emptyset$, then μ is dominated by μ^*, which differs from μ only in that the types which prefer ν switch to ν. If ν is incentive compatible given S then μ^* is incentive compatible (since μ is incentive compatible given $T_1 \setminus S$); but this contradicts the assumption that μ is undominated.

To prove the last sentence of the theorem, let $\nu = \hat{\mu}$. Since $\hat{\mu}$ is incentive compatible given any set, $\{t_1 \mid U_1(\hat{\mu}|t_1) > U_1(\mu|t_1)\} = \emptyset$. Similarly, switching the roles of $\hat{\mu}$ and μ, we get
$\{t_1 \mid U_1(\mu|t_1) > U_1(\hat{\mu}|t_1)\} = \emptyset$. Q.E.D.

Theorem 1 shows us why the principal should implement a strong solution (if one exists), even though he might actually (given his true type) prefer some other incentive-compatible mechanism. If the subordinates were to interpret his selection of any other mechanism ν as evidence that his type must be in the set preferring ν over the strong solution, then ν would become infeasible as soon as it was selected. Furthermore, Theorem 1 states that, if a strong solution exists, it must be essentially unique.

Let us now consider Example 1, an incentive problem with one subordinate (so $n = 2$). The principal has two equally-likely types $T_1 = \{1a, 1b\}$. The subordinate has no private information, so $|T_2| = 1$ and the variable t_2 can be ignored. There are three enforceable actions in $D_0 = \{\sigma_0, \sigma_1, \sigma_2\}$ available to the principal, but he has no private options (so $|D_1| = 1$ and the variable d_1 can be ignored). The subordinate has two private actions $D_2 = \{\bar{r}, \bar{a}\}$. If the subordinate chooses $\bar{r}$ ("reject") then both individuals will get payoffs of zero. If the subordinate chooses $\bar{a}$ ("accept") then the individuals' utility payoffs (u_1,u_2) depend on d_0 and t_1 as in the following table:

	$d_0=\sigma_0$	$d_0=\sigma_1$	$d_0=\sigma_2$
t_1=1a	0,0	9,-2	5,3
t_1=1b	0,0	5,3	9,-2

TABLE 1.

In this example (by (2.7) and (2.8)), an incentive-compatible mechanism must give nonnegative expected utility to the subordinate, and must not give either type of the principal any incentive to report the other type. Among such mechanisms, the expected utility for type 1a is maximized by the mechanism μ_1, defined by

$$\mu_1(\sigma_1,\bar{a}|1a) = \mu_1(\sigma_1,\bar{a}|1b) = 1.$$

The expected utility for type 1b is maximized by the mechanism μ_2, defined by

$$\mu_2(\sigma_2,\bar{a}|1a) = \mu_2(\sigma_2,\bar{a}|1b) = 1.$$

That is, type 1a prefers the mechanism μ_1, in which both of the principal's types implement σ_1; and type 1b prefers the mechanism μ_2, in which both types implement σ_2. These mechanisms give expected utilities as follows:

$$U_1(\mu_1|1a) = 9,\quad U_1(\mu_1|1b) = 5,\quad U_2(\mu_1) = .5,$$

$$U_1(\mu_2|1a) = 5,\quad U_1(\mu_2|1b) = 9,\quad U_2(\mu_2) = .5$$

Unfortunately for the principal, neither μ_1 nor μ_2 is incentive compatible unless both types are expected to implement it. If the principal were expected to choose μ_1 only if his type is 1a, and to choose μ_2 only if his type is 1b, then the subordinate's expected utility would be -2 in each case and he would be better off rejecting. Although the subordinate is willing to accept either σ_1 or σ_2 ex ante, he would prefer to reject against either action if he knew that it was the one that the principal preferred.

To guarantee that the subordinate would be willing to accept, no matter what he might infer, the principal could offer to randomize between σ_1 and σ_2. For example he could use μ_3, defined by

$$\mu_3(\sigma_1,\bar{a}|t_1) = \mu_3(\sigma_2,\bar{a}|t_1) = .5,\quad \forall t_1 \varepsilon \{1a, 1b\}.$$

This mechanism μ_3 is safe, since the subordinate's expected utility would be nonnegative (+.5) even if he learned the principal's type. However μ_3 is not undominated. The unique safe and undominated mechanism is μ_4, defined by

$$\mu_4(\sigma_1,\bar{a}|1a) = .6,\quad \mu_4(\sigma_2,\bar{a}|1a) = .4,$$

$$\mu_4(\sigma_1,\bar{a}|1b) = .4,\quad \mu_4(\sigma_2,\bar{a}|1b) = .6$$

That is, μ_4 is the mechanism in which the subordinate never rejects, and the principal randomizes between σ_1 and σ_2, giving at most 60% probability to the action that he actually prefers. (The trustworthy mediator, described in Section 2, could verify to the subordinate that the randomization was actually carried out within these .60 - .40 bounds.) The subordinate expects zero utility from μ_4, with either type of the principal, and the principal's expected-utility allocation is $U_1(\mu_4|1a) = 7.4 = U_1(\mu_4|1b)$. Thus, although μ_4 is not the best incentive-compatible mechanism for either type, it is the principal's strong solution. Any mechanism ν that offers higher expected utility to either type of the principal would be rejected by the subordinate, because he would expect negative utility from the mechanism after inferring that the principal was of the type for which $U_1(\nu|t_1) > U_1(\mu_4|t_1)$.

5. Expectational Equilibria

In the preceding section, we argued that, if there exists a mechanism that is both safe and undominated, then this (essentially unique) mechanism should be implemented by all types of the principal. To fully justify this claim, and to begin to derive a theory of rational selection of a mechanism by the principal for the general case in which a strong solution may not exist, we must consider the principal's selection of a mechanism as part of a noncooperative game.

In this noncooperative game, each individual first learns his own type; then the principal selects and announces a coordination mechanism; then each subordinate makes some inferences about the principal's type, based on this announcement; and finally the coordination mechanism is implemented, with each individual using some participation strategy that is rational for him given his information. To rigorously analyze this game, we must first develop some notation.

For any vector q in $\mathbb{R}^{T_1}$ such that $0 \leq q(t_1) \leq 1 \ \forall t_1$, and $q \neq \underset{\sim}{0}$, we let

(5.1) $$p_i^*(t_{-i}|t_i,q) = p_i(t_{-i}|t_i)\, q(t_1)/\Big(\sum_{s_{-i}\varepsilon T_{-i}} p_i(s_{-i}|t_i)\, q(s_1)\Big),$$

$$\forall i\varepsilon\{2,\ldots,n\},\ \forall t\varepsilon T,$$

and

$$p_1^*(t_{-1}|t_1,q) = p_1(t_{-1}|t_1),\quad \forall t\varepsilon T.$$

To interpret this definition, suppose that, for each t_1, $q(t_1)$ is the likelihood (or conditional probability) of the principal selecting mechanism ν when his type is t_1. Then $p_i^*(t_{-i}|t_i,q)$ is the posterior probability that individual i would assign to state t if his own type were t_i and the principal selected mechanism ν.

For any likelihood vector q as above, the normalized-likelihood vector Q corresponding to q is defined by

(5.2) $$Q(t_1)\Big(\sum_{s_1\varepsilon T_1} q(s_1)\Big) = q(t_1),\quad \forall t_1\varepsilon T_1.$$

Notice that (5.2) implies that $p_i^*(t_{-i}|t_i,q) = p_i^*(t_{-i}|t_i,Q)$, for every i and t, for any nonzero vector q. Thus, we need to know only the normalized-likelihood vector Q associated with a mechanism ν, to compute the individuals' posterior probabilities if ν were selected by the principal.

Suppose now that mechanism ν has zero likelihood of being selected by each type of the principal, so that $q = \tilde{0}$. Then the posterior probabilities if ν actually were selected cannot be computed from (5.1), because the denomimator of (5.1) is zero. (The regularity assumption (2.1) prevented this difficulty in all other cases.) On the other hand, (5.2) is satisfied by any normalized-likelihood vector Q when $q = \tilde{0}$. Thus, following Kreps and Wilson [1982], we may say that the individuals' posterior beliefs after the selection of ν are consistent iff there exists some normalized-likelihood vector Q, satisfying

$$(5.3) \qquad \sum_{s_1 \varepsilon T_1} Q(s_1) = 1 \quad \text{and} \quad Q(t_1) \geq 0, \quad \forall t_1 \varepsilon T_1,$$

such that, for every i and t, $p_i^*(t_{-i}|t_i,Q)$ is the posterior probability that individual i would assign to state t if his own type were t_i and the principal selected ν. This vector Q may be interpreted as the normalized-likelihood vector corresponding to some vector of nonzero but infinitesimal likelihoods of the principal selecting ν.

We do not need to assume that the principal must select a direct revelation mechanism. A generalized mechanism is defined to be any function $\nu:D'\times T' \to \mathbb{R}$ such that D' and T' are nonempty finite sets of the form

$$D' = D_0 \times D_1' \times \ldots \times D_n', \qquad T' = T_1' \times \ldots \times T_n',$$

and

$$\sum_{c \varepsilon D'} \nu(c|t) = 1 \quad \text{and} \quad \nu(d|t) \geq 0, \quad \forall d \varepsilon D', \quad \forall t \varepsilon T'.$$

Here T_i' is the set of possible reports that i may send, D_i' is the set of possible instructions that i may receive, and $\nu(d|t)$ is the probability of implementing d_0 and sending instructions d_i to each i, if each i has reported t_i into the mechanism. The only change from (2.3) is that D_i' and T_i' may differ from D_i and T_i.

When the generalized mechanism ν is implemented, each individual i will determine his reports and private actions according to some participation strategy, denoted by a pair (γ_i,τ_i) such that

$$(5.4) \qquad \sum_{s_i \varepsilon T_i'} \tau_i(s_i|t_i) = 1 \quad \text{and} \quad \tau_i(r_i|t_i) \geq 0, \quad \forall r_i \varepsilon T_i', \ \forall t_i \varepsilon T_i;$$

and

$$(5.5)\qquad \sum_{c_i \varepsilon D_i} \gamma_i(c_i|d_i,s_i,t_i) = 1 \quad\text{and}\quad \gamma_i(b_i|d_i,s_i,t_i) \geq 0,$$

$$\forall b_i \varepsilon D_i,\ \forall d_i \varepsilon D_i',\ \forall s_i \varepsilon T_i',\ \forall t_i \varepsilon T_i.$$

Here $\tau_i(s_i|t_i)$ is the probability that i will report s_i if his type is t_i; and $\gamma_i(c_i|d_i,s_i,t_i)$ is the probability that i will use his private action c_i if t_i is his true type but he reported s_i and then received instructions d_i, in the implementation of ν.

We let $W_i(\nu,\gamma,\tau|t_i,Q)$ denote the expected utility for individual i in the mechanism ν if his type is t_i, his posterior distribution given the selection of ν is characterized by the normalized-likelihood vector Q, and $\gamma = (\gamma_1,\ldots,\gamma_n)$ and $\tau = (\tau_1,\ldots,\tau_n)$ characterize the participation strategies of the n individuals. That is

$$(5.6)\qquad W_i(\nu,\gamma,\tau|t_i,Q) =$$

$$= \sum_{t_{-i}\varepsilon T_{-i}} \ \sum_{s\varepsilon T'} \ \sum_{d\varepsilon D'} \ \sum_{c\varepsilon D} p_i^*(t_{-i}|t_i,Q)\ \tau(s|t)\ \nu(d|s)\ \gamma(c|d,s,t)\ u_i(c,t)$$

where

$$\tau(s|t) = \prod_{j=1}^{n} \tau_j(s_j|t_j)$$

and

$$\gamma(c|d,s,t) = \begin{cases} \prod_{j=1}^{n} \gamma_j(c_j|d_j,s_j,t_j) & \text{if } c_0 = d_0, \\ 0 & \text{if } c_0 \neq d_0. \end{cases}$$

We say that the participation strategies (γ,τ) are a <u>Nash equilibrium for ν given Q</u> iff every individual's participation strategy maximizes the expected utility for each type, given the other individuals' strategies, so that

$$W_i(\nu,\gamma,\tau|t_i,Q) \geq W_i(\nu,(\gamma_{-i},\gamma_i'),\ (\tau_{-i},\tau_i')|t_i,Q)$$

for every i in $\{1,\ldots,n\}$, every t_i in T_i, and every alternative participation strategy (τ_i', γ_i') satisfying (5.4) and (5.5) for i.

We say that a mechanism μ is an <u>expectational equilibrium</u> iff μ is incentive compatible and, for every generalized mechanism ν, there exist Q, γ, and τ satisfying (5.3)-(5.5) such that (γ,τ) is a Nash equilibrium for ν given Q and

$$(5.7) \qquad U_1(\mu|t_1) \geq W_1(\nu,\gamma,\tau|t_1,Q), \quad \forall t_1 \varepsilon T_1.$$

In the terminology of Kreps and Wilson [1982], any expectational equilibrium can be supported as a <u>sequential equilibrium</u> of the mechanism-selection game.

When μ is an expectational equilibrium then rational behavior of the subordinates can force all types of the principal to implement μ. If he were to try to implement some other mechanism ν then, with the posterior expectations characterized by Q, the subordinates would find it rational to use their equilibrium participation strategies (γ,τ). By (5.7), these participation strategies in ν would leave the principal no better off than in μ, no matter what his type may be. So all of the principal's types would prefer to implement μ. But then any posterior probabilities characterized by any normalized likelihood vector such as Q would be consistent with rational Bayesian inference after the principal selected ν, because the event of ν being selected has zero probability for every type in T_1.

Theorem 1 did not completely justify our claim that, if μ is a strong solution, then all types of the principal should select it. Theorem 1 showed that any alternative mechanism ν could not be incentive compatible given the information that the principal would prefer it, implemented honestly and obediently, over μ. One could still ask, however, whether the subordinates would use some dishonest or disobedient strategies in μ such that some types of the principal would be better off than in μ. The following theorem shows that the answer to this question is No, because for any alternative mechanism there are consistent posterior beliefs and a Nash equilibrium of participation strategies such that no type of the principal would be better off than in the strong solution.

Theorem 2. Any strong solution is an expectational equilibrium.

We defer the proof of this theorem to Section 9.

The concept of expectational equilibrium can be applied to any Bayesian incentive problem, even if there is no strong solution. We prove the following general existence theorem in Section 9.

Theorem 3 For any Bayesian incentive problem as in (2.2), there exists at least one expectational equilibrium.

6. Core Mechanisms

For Example 1 (discussed in Section 4), one can show that the strong solution μ_4 is the unique expectational equilibrium. Unfortunately, expectational equilibria are not generally unique (even when a strong solution exists), and for some Bayesian incentive problems the set of expectational equilibria may be quite large. Thus, to get a more useful theory, we must investigate other solution concepts for the informed principal's problem.

Let us consider now Example 2, which differs from Example 1 only in that the utility functions (u_1,u_2) are as follows:

	$d_0=\sigma_0$	$d_0=\sigma_1$	$d_0=\sigma_2$
t_1=1a	0,0	9,-2	5,-1
t_1=1b	0,0	5,3	9,1

TABLE 2

(The only changes are in the subordinate's payoffs from outcome σ_2.) As before, the subordinate (individual 2) believes ex ante that types 1a and 1b are equally likely, and he has the option to reject the principal's mechanism, in which case both individuals' payoffs are zero. Thus, a mechanism is incentive compatible iff it gives nonnegative expected utility to the subordinate, and gives neither type of the principal any incentive to lie.

In this example, every incentive-compatible mechanism is an expectational equilibrium (including even the strictly dominated mechanisms). This is because the subordinate would choose to reject against any mechanism if he believed that the principal was type 1a. Thus, to show that any given incentive-compatible mechanism μ is an expectational equilibrium, let Q(1a) = 1 and Q(1b) = 0 for any alternative mechanism ν. Then both types of the principal should rationally select μ, because the subordinate believes that only type 1a could make the "mistake" of choosing anything else, and so he would reject anything else.

As before, let μ_1 be the mechanism in which both of the principal's types do σ_1, and let μ_2 be the mechanism in which both types do σ_2. Although each of these mechanisms is an expectational equilibrium, there is good reason to believe that the principal should actually implement μ_2. After all, μ_2 is the best incentive-compatible mechanism for type 1b. Thus, it would seem strange for the subordinate to infer that the principal is type 1a when μ_2 is announced, as was required to make μ_1 an expectational equilibrium.

The weakness of expectational equilibrium as a solution concept is that it allows so much flexibility in the designation of the posterior beliefs after an alternative mechanism is selected. If the principal is able to communicate effectively with his subordinates, he may actually have some control over the subordinates posterior beliefs, to the extent that he can explain why he is choosing a particular mechanism. In Example 2, if the subordinate were expecting both types of the principal to use μ_1, then the principal could speak to the subordinate as follows:

> "I am going to implement μ_2. This mechanism is strictly better for 1b, but worse for 1a, than the mechanism μ_1 which you may have been expecting me to implement. Thus you should take my selection of μ_2 as evidence in favor of my being type 1b. But whether you infer that my type is 1b, or you remain with your prior belief that my two types are equally likely (or even if you make any inference in between), this mechanism μ_2 gives you nonnegative expected utility. Thus, you should not reject against μ_2."

If the principal can communicate effectively, then the subordinate should understand this speech and accept it. Furthermore, if it is common knowledge that he would accept it, then the subordinate could not rationally expect type 1b to choose μ_1, so he should reject against μ_1.

The above argument can be extended to the general case. We say that μ is a <u>core mechanism</u> for the principal iff μ is incentive compatible and there does not exist any other mechanism ν such that

$$\{t_1 \varepsilon T_1 \mid U_1(\nu|t_1) > U_1(\mu|t_1)\} \neq \emptyset$$

and, for every set S that satisfies

$$\{t_1 \varepsilon T_1 \mid U_1(\nu|t_1) > U_1(\mu|t_1)\} \subseteq S \subseteq T_1,$$

ν would be incentive-compatible given S. That is, if μ is not a core mechanism, then there is some other mechanism ν that some types would prefer, such that ν would be incentive compatible given the information revealed by its selection, provided that (at least) all the types that prefer ν over μ are expected to choose ν. In Example 2, μ_2 is the unique core mechanism. The following existence theorem is proven in Section 9.

Theorem 4. For any Bayesian incentive problem, there exists at least one core mechanism for the principal.

The term "core mechanism" suggests a connection with cooperative game theory. Indeed, these core mechanisms can be characterized as the core of a cooperative game, but one in which the players are different types of the principal, rather than different individuals. In this cooperative game, the set of feasible mechanisms for a coalition S is the set of mechanisms that would be incentive compatible given any superset of S.

It is not surprising that the problem of mechanism design by informed principal should have parallels with cooperative game theory. The principal's problem is to select a mechanism that will be perceived as a reasonable compromise between the different goals of his different possible types, and the problem of reasonable compromise between conflicting goals of different individuals is the subject of cooperative game theory.

It also should not be surprising that the methods of noncooperative game theory, as embodied in the concept of expectational equilibrium, are not generally sufficient to determine the solution to the informed principal's problem. Our notion of principal is meant to refer to someone who can communicate effectively with his subordinates in some common language like

English. We used this assumption of effective communication when we quoted hypothetical speeches that a principal could make to justify his selection of an undominated or core mechanism. However, noncooperative game theory is meant to apply also to situations in which the individuals might not share any common language. Thus, our assumption of effective communication has required us to go beyond the scope of existing noncooperative game theory.

7. Blocked Allocations and Neutral Mechanisms

Thus far, we have developed a variety of solution concepts for the principal's problem: undominated mechanisms, strong solutions, expectational equilibria, and core mechanisms. A strong solution is essentially unique when it exists, but it may fail to exist. Mechanisms that satisfy the other three solution concepts can be shown to always exist, but the set of such mechanisms may be quite large. (There was a unique core mechanism in each of the two examples above, but other examples can be constructed in which the entire continuum of undominated mechanisms are core mechanisms.) We still want a more powerful solution concept, to identify the principal's best inscrutable mechanisms.

For inscrutability, the informed principal must select a mechanism that would seem like a reasonable selection for all of his types to make. Some mechanisms would clearly be unreasonable selections for some types, when these types could do better by selecting some other mechanism (even though they may reveal some information by doing so). That is, some mechanisms may be blocked or eliminated for the principal because they give too low expected utility to some types of the principal.

For any mechanism μ, we let $U_1(\mu)$ denote the expected utility allocation vector for the principal's types; that is

$$U_1(\mu) = (U_1(\mu|t_1))_{t_1 \varepsilon T_1} \varepsilon \mathbb{R}^{T_1} .$$

By the above argument, some mechanisms may be blocked for the principal because some components of this allocation vector are too low. Thus, for any Bayesian incentive problem Γ (as in (2.2)), there should be some set $B(\Gamma)$, a subset of $\mathbb{R}^{T_1}$, such that $B(\Gamma)$ represents the set of blocked allocation vectors. Our theoretical task is to determine what this set $B(\Gamma)$ should be, for every Bayesian incentive problem Γ. Then we could say that an incentive-compatible mechanism μ would be a reasonable selection for all types of the principal, in the incentive problem Γ, only if the allocation $U_1(\mu)$ is not in the blocked set $B(\Gamma)$.

Most of the solution concepts that we have discussed so far can be characterized in terms of such sets of blocked allocations, with a different $B(\Gamma)$ for each solution concept. Let us now approach the problem of constructing a new solution concept more systematically. We list four properties that the sets of "blocked" allocations should satisfy, and then construct the largest $B(\Gamma)$ sets that can satisfy these four axioms.

Our first axiom expresses the idea that an allocation vector is blocked when some of its components are too low. Thus, any other vector that is componentwise lower than a blocked allocation vector should also be blocked.

Axiom 1. (Domination) For any Bayesian incentive problem Γ and any vectors w and x in $\mathbb{R}^{T_1}$, if $w \varepsilon B(\Gamma)$ and $x(t_1) \leqslant w(t_1)$ for every t_1 in T_1 then $x \varepsilon B(\Gamma)$.

If the blocking of an allocation w is supposed to occur because some types could do strictly better by selecting some other mechanism, then these strict inequalities would also hold for all allocation vectors that are sufficiently close to w. Thus, $B(\Gamma)$ should be an open set.

Axiom 2. (Openness) For any incentive problem Γ, $B(\Gamma)$ is an open subset of $\mathbb{R}^{T_1}$.

Consider any two Bayesian incentive problems Γ and $\bar{\Gamma}$ where

$$\Gamma = (D_0, D_1, \ldots, D_n, T_1, \ldots, T_n, u_1, \ldots, u_n, p_1, \ldots, p_n)$$

and

$$\bar{\Gamma} = (\bar{D}_0, \bar{D}_1, \ldots, \bar{D}_n, \bar{T}_1, \ldots, \bar{T}_n, \bar{u}_1, \ldots, \bar{u}_n, \bar{p}_1, \ldots \bar{p}_n).$$

We say that $\bar{\Gamma}$ is an extension of Γ iff:

$$\bar{D}_i = D_i,\ \bar{T}_i = T_i,\ \bar{p}_i = p_i,\quad \forall i \varepsilon \{1, \ldots, n\};$$

$$\bar{D}_0 \supseteq D_0; \text{ and}$$

$$\bar{u}_i(d,t) = u_i(d,t) \text{ whenever } d_0 \varepsilon D_0.$$

That is, $\bar{\Gamma}$ is an extension of Γ iff $\bar{\Gamma}$ differs from Γ only in that some new enforceable actions have been added to those in D_0. Every incentive-compatible mechanism available to the principal in Γ is also available in any extension $\bar{\Gamma}$. Thus, the set of blocked allocations in any extension of Γ should be at least as large as in Γ, because there are more mechanisms available with which the types in T_1 can block.

Axiom 3. (Extensions) If $\bar{\Gamma}$ is any extension of an incentive problem Γ, then $B(\bar{\Gamma}) \supseteq B(\Gamma)$.

We have argued (by Theorems 1 and 2) that, if there exists a mechanism that is both safe and undominated, then this mechanism can be called a strong solution and all types of the principal should select it or some other mechanism that gives the same utility allocation. Thus, these strong solutions must not be blocked.

Axiom 4. (Strong Solutions) If μ is a safe and undominated mechanism for the principal in an incentive problem Γ, then $U_1(\mu) \notin B(\Gamma)$.

We let $\underset{\sim}{H}$ denote the set of all functions $B(\cdot)$ (mapping Bayesian incentive problems into subsets of the principal's utility-allocation space) that satisfy all four of these axioms, and we let

$$B^*(\Gamma) = \bigcup_{B \in \underset{\sim}{H}} B(\Gamma)$$

for any incentive problem Γ. That is, $B^*(\Gamma)$ is the union of all sets of allocations that can be blocked in Γ, consistently with Axioms 1 through 4. It is straightforward to check that $B^*(\cdot)$ itself satisfies Axioms 1 through 4.

Given any Bayesian incentive problem Γ, we say that μ is a neutral optimum for the principal iff μ is an incentive-compatible mechanism and $U_1(\mu)$ is not in $B^*(\Gamma)$. That is, a neutral optimum is an incentive-compatible mechanism that cannot be blocked by any theory of "blocking" that satisfies our four axioms. Thus, the neutral optima form the smallest class of mechanisms that we could hope to identify as solutions for the principal.

It is shown in Section 9 that expectational equilibria and core mechanisms can both be characterized as sets of unblocked incentive-compatible mechanisms, in terms of some blocking concepts that satisfy the four axioms. By Axiom 4, strong solutions are never blocked. Thus, we get the following theorem.

Theorem 5. Any safe and undominated mechanism is a neutral optimum. Any neutral optimum is both an expectational equilibrium and a core mechanism.

In Section 9 we also prove our main existence theorem, from which Theorems 3 and 4 will follow immediately.

Theorem 6. For any Bayesian incentive problem, there exists at least one neutral optimum for the principal.

From these two theorems, we can determine the principal's neutral optima in our two examples. In Example 1, μ_4 was the unique expectational equilibrium, so it must also be the unique neutral optimum. In Example 2, μ_2 was the unique core mechanism, although there were infinitely many expectational equilibria; so μ_2 is the unique neutral optimum in Example 2.

8. Characterizing the Neutral Optima

Given any Bayesian incentive problem Γ as in (2.2), we now show how to characterize the set of neutral optima for the principal. First, some notation should be developed.

We let Ω denote the set of vectors or functions on T_1 into the real numbers $\mathbb{R}$; that is,

$$\Omega = \mathbb{R}^{T_1}.$$

We let Ω_+ denote the set of nonnegative-valued functions in Ω, and we let Ω_{++} denote the set of strictly positive-valued functions in Ω. That is,

$\lambda\varepsilon\Omega_+$ iff $\lambda(t_1) \geq 0$ for every t_1 in T_1; and $\lambda\varepsilon\Omega_{++}$ iff $\lambda(t_1) > 0$ for every t_1 in T_1.

The set of incentive-compatible mechanisms is a convex polyhedron (since D and T are finite). Thus, by the supporting hyperplane theorem, a mechanism μ^* is undominated if and only if there is some λ in Ω_{++} such that μ^* is an optimal solution to the problem

$$(8.1) \qquad \underset{\mu}{\text{maximize}} \sum_{t_1 \varepsilon T_1} \lambda(t_1)\, U_1(\mu|t_1), \quad \text{subject to (2.3) and (2.6).}$$

We may refer to (8.1) as the primal problem for λ. With finite D and T, it is a linear programming problem.

To formulate its dual, let Δ_i denote the set of all functions from D_i into D_i. We define

$$(8.2) \qquad A = \{\alpha\varepsilon \mathop{\times}_{i=1}^{n} \mathbb{R}^{\Delta_i \times T_i \times T_i} \mid \alpha_i(\delta_i, s_i|t_i) \geq 0,\ \forall i,\ \forall \delta_i \varepsilon \Delta_i,\ \forall s_i \varepsilon T_i,\ \forall t_i \varepsilon T_i\}.$$

For any α in A, we will interpret $\alpha_i(\delta_i, s_i|t_i)$ as a shadow price for the primal constraint $U_i(\mu|t_i) \geq U_i^*(\mu, \delta_i, s_i|t_i)$.

For any d in D, t in T, λ in Ω_+, and α in A, we define

$$(8.3) \qquad L(d,t,\lambda,\alpha) = \Big(\lambda(t_1)\, p_1(t_{-1}|t_1)\, u_1(d,t)$$

$$+ \sum_{i=1}^{n} \sum_{s_i \varepsilon T_i} \sum_{\delta_i \varepsilon \Delta_i} \alpha_i(\delta_i, s_i|t_i)\, p_i(t_{-i}|t_i)\, u_i(d,t)$$

$$- \sum_{i=1}^{n} \sum_{s_i \varepsilon T_i} \sum_{\delta_i \varepsilon \Delta_i} \alpha_i(\delta_i, t_i|s_i)\, p_i(t_{-i}|s_i)\, u_i((d_{-i}, \delta_i(d_i)), (t_{-i}, s_i))\Big).$$

When the incentive constraints (2.6) are multiplied by their shadow prices and added into the primal objective function, we get the Lagrangian function

$$(8.4)\qquad \Big(\sum_{t_1\varepsilon T_1} \lambda(t_1)\, U_1(\mu|t_1)$$

$$+ \sum_{i=1}^{n} \sum_{t_1\varepsilon T_i} \sum_{s_i\varepsilon T_i} \sum_{\delta_i\varepsilon\Delta_i} \alpha_i(\delta_i,s_i|t_i)\,(U_i(\mu|t_i)-U_i^*(\mu,\delta_i,s_i|t_i))\Big)$$

$$= \sum_{t\varepsilon T} \sum_{d\varepsilon D} L(d,t,\lambda,\alpha)\, \mu(d|t).$$

That is, $L(d,t,\lambda,\alpha)$ has been defined as the linear coefficient of the term $\mu(d|t)$ in the Lagrangian function. The Lagrangian is maximized by μ (subject to the remaining probability constraints (2.3)) iff all probability weight in each $\mu(\cdot|t)$ distribution is put on the outcomes d that maximize $L(d,t,\lambda,\alpha)$. Thus, the <u>dual problem</u> for λ (the dual to (8.1)) may be written as

$$(8.5)\qquad \underset{\alpha\varepsilon A}{\text{minimize}} \sum_{t\varepsilon T} \max_{d\varepsilon D} L(d,t,\lambda,\alpha).$$

When we vary λ as a free parameter over Ω_{++}, the optimal solutions to the primal (8.1) cover the entire set of undominated mechanisms for the principal. Our problem is to characterize which of these mechanisms are neutral optima for the principal.

One more bit of notation will be useful. For any α in A, we define

$$(8.6)\qquad \alpha_i(s_i|t_i) = \sum_{\delta_i\varepsilon\Delta_i} \alpha_i(\delta_i,s_i|t_i).$$

That is, $\alpha_i(s_i|t_i)$ is the aggregated shadow price for the constraint that type t_i of individual i should not be tempted to claim to be type s_i.

We can now state the necessary and sufficient conditions that characterize the principal's neutral optima.

Theorem 7. An incentive-compatible mechanism μ is a neutral optimum for the principal if and only if there exist sequences $\{\lambda^k, \alpha^k, \omega^k\}_{k=1}^{\infty}$ such that

$$\text{(8.7)} \qquad \lambda^k \varepsilon \Omega_{++},\ \alpha^k \varepsilon A,\ \omega^k \varepsilon \Omega,\ \forall k;$$

$$\text{(8.8)} \qquad \Big(\lambda^k(t_1) + \sum_{s_1 \varepsilon T_1} \alpha_1^k(s_1|t_1)\Big)\omega^k(t_1) - \sum_{s_1 \varepsilon T_1} \alpha_1^k(t_1|s_1)\omega^k(s_1)$$

$$= \sum_{t_{-1} \varepsilon T_{-1}} \max_{d \varepsilon D} L(d,t,\lambda^k,\alpha^k), \quad \forall t_1 \varepsilon T_1, \quad \forall k;$$

$$\text{(8.9)} \qquad \limsup_{k \to \infty} \omega^k(t_1) \leqslant U_1(\mu|t_1), \quad \forall t_1 \varepsilon T_1.$$

This theorem is proved in Section 9.

To interpret Theorem 7, one must understand equation (8.8). We say that a vector ω in Ω is warranted by λ and α (and $\omega(t_1)$ is the warranted claim of type t_1) iff

$$\text{(8.10)} \qquad \Big(\lambda(t_1) + \sum_{s_1 \varepsilon T_1} \alpha_1(s_1|t_1)\Big)\ \omega(t_1) - \sum_{s_1 \varepsilon T_1} \alpha(t_1|s_1)\ \omega(s_1)$$

$$= \sum_{t_{-1} \varepsilon T_{-1}} \max_{d \varepsilon D} L(d,t,\lambda,\alpha), \quad \forall t_1 \varepsilon T_1.$$

Lemma 2 in Section 9 asserts that, if ω is warranted by some λ in Ω_{++} and α in A, then there exists an extension of Γ in which a strong solution gives, to each type t_1 of the principal, an expected utility equal to $\omega(t_1)$. Theorem 7 asserts that μ is a neutral optimum if and only if there are such warranted utility allocations for the principal in which no type's warranted claim exceeds what it gets from μ by more than an arbitrarily small amount.

The following theorem lists some simpler necessary conditions for a neutral optimum.

Theorem 8. If μ is a neutral optimum then there exist λ in Ω_+, α in A, and ω in Ω, such that

(8.11) μ is an optimal solution of the primal problem for λ,

(8.12) α is an optimal solution of the dual problem for λ,

(8.13) ω is warranted by λ and α,

(8.14) $\omega(t_1) \leq U_1(\mu|t_1)$ and $\lambda_1(t_1)\big(\omega(t_1) - U_1(\mu|t_1)\big) = 0, \quad \forall t_1 \varepsilon T_1$.

(8.15) $(\lambda,\alpha) \neq (\underset{\sim}{0},\underset{\sim}{0})$.

The proof is deferred to Section 9.

Notice that the conditions of Theorem 8 form a well-determined system, in the sense of having as many equations as variables. Condition (8.15) is a nontriviality condition, requiring that at least one component of λ or α must be strictly positive. By (8.11)-(8.13), the primal problem (8.1) determines μ, the dual problem (8.5) determines α, and the warrant equations (8.10) determine ω. Finally, (8.14) gives us as many equations $(\omega(t_1) = U_1(\mu|t_1)$ or $\lambda(t_1) = 0)$ as there are parameters $\lambda(t_1)$ to be determined. This suggests a conjecture that the set of neutral optima may be generically finite.

Examples are known in which there are several neutral optima for the principal. By the axiomatic definition of neutral optima, no game-theoretic concept of blocking that satisfies our four axioms can eliminate any of these neutral optima. Extra-game-theoretic considerations of history or tradition may be decisive in determining the principal's selection of a mechanism when there are many neutral optima.

Let us consider the special case mentioned in Section 2, in which the general incentive constraints (2.6) can be replaced by simpler self-selection constraints (2.7) and nonnegative-payoff constraints (2.8). In this case, the Lagrangian coefficients can be written

$$L(d,t,\lambda,\alpha) = \Big(\lambda(t_1)\ p_1(t_{-1}|t_1)\ u_1(d,t) + \sum_{i=2}^{n} \alpha_{i,0}(t_i)\ p_i(t_{-i}|t_i)\ u_i(d,t)$$

$$+ \sum_{i=1}^{n} \sum_{s_i \varepsilon T_i} \alpha_i(s_i|t_i)\ p_i(t_{-i}|t_i)\ u_i(d,t)$$

$$- \sum_{i=1} \sum_{s_i \varepsilon T_i} \alpha_i(t_i|s_i)\ p_i(t_{-i}|s_i)\ u_i(d,(t_{-i},s_i))\Big),$$

where $\alpha_{i,0}(t_i)$ is the dual variable for the constraint $U_i(\mu|t_i) \geqslant 0$, and $\alpha_i(s_i|t_i)$ is the shadow price of the constraint that type t_i should not expect to gain by reporting type s_i. With this one modification, Theorems 7 and 8 can be adapted to this case.

For example, consider Example 2, in which we saw that μ_2 must be the unique neutral optimum for the principal. To verify the conditions of Theorem 7, let

$$\lambda^k(1a) = 1/k,\quad \lambda^k(1b) = 1,\quad \alpha_1^k(1a|1b) = \alpha_1^k(1b|1a) = 0,\quad \alpha_{2,0}^k = 5/k.$$

Then for any $k \geqslant 3$, the warranted claims satisfying (8.8) are $\omega^k(1a) = 0$ and $\omega^k(1b) = 9 + 5/k$, so

$$\lim_{k\to\infty} \omega^k(1a) = 0 < 5 = U_1(\mu_2|1a),$$

$$\lim_{k\to\infty} \omega^k(1b) = 9 = U_1(\mu_2|1b).$$

9. Proofs

We prove the theorems in the following order: 2, 7, 8, 6, 5, 3 and 4.

Proof of Theorem 2. Let μ be a safe and undominated mechanism in the incentive problem Γ. We must show that μ is an expectational equilibrium.

Let ν be any generalized mechanism, and consider a mechanism-selection game in which the principal must select either μ or ν. Since μ is incentive compatible given any of the principal's types, we may assume that the individuals will participate honestly and obediently in μ if he selects it. Let $q(t_1)$ denote the probability that the principal selects ν if his type is t_1; let Q be a normalized likelihood vector corresponding to q, satisfying (5.2); and let (γ,τ) denote the participation strategies that the individuals would use in ν. In a sequential equilibrium of this mechanism-selection game, we must have

$$(9.1) \qquad q(t_1) = \begin{cases} 1 \text{ if } W_1(\nu,\gamma,\tau|t_1,Q) > U_1(\mu|t_1), \\ 0 \text{ if } W_1(\nu,\gamma,\tau|t_1,Q) < U_1(\mu|t_1), \end{cases}$$

and (γ,τ) must be a Nash equilibrium of ν given Q. Condition (5.2) uniquely determines Q, unless $q = \underset{\sim}{0}$, in which case any Q in the unit simplex will do. By a straightforward argument using the Kakutani fixed point theorem, it can be shown that such a sequential equilibrium (q,Q,γ,τ) does exist.

This equilibrium of the mechanism-selection game is equivalent to the direct revelation mechanism η, defined by

$$\eta(d|t) = q(t_1)\Big(\sum_{s\in T'}\sum_{c\in D'} \tau(s|t)\,\nu(c|s)\,\gamma(d|t)\Big) + \big(1-q(t_1)\big)\,\mu(d|t).$$

By (9.1), $U_1(\eta|t_1) = \max\{U_1(\mu|t_1), W_1(\nu,\gamma,\tau|t_1,Q)\}$. Furthermore, η is incentive compatible, because μ is safe and the individuals are using equilibrium participation strategies in ν given their rational beliefs following its selection. But μ is not dominated by any incentive-compatible mechanism, so $U_1(\mu|t_1) \geq W_1(\nu,\gamma,\tau|t_1,Q)$ for every t_1. Thus (γ,τ,Q) support μ as an expectational equilibrium over ν. Q.E.D.

We now state and prove three lemmas.

Lemma 1. Given any α in A, h in Ω, and λ in Ω_{++}, there is a unique vector ω in Ω such that

$$(9.2)\qquad \Big(\lambda(t_1) + \sum_{s_1\in T_1} \alpha_1(s_1|t_1)\Big)\omega(t_1) - \sum_{s_1\in T_1} \alpha_1(t_1|s_1)\,\omega(s_1) = h(t_1),\quad \forall t_1\in T_1.$$

Furthermore, the solution ω to these linear equations is increasing in the vector h. (That is, if $h'(t_1) \geq h(t_1)$ $\forall t_1$, and ω' solves (9.2) for h' instead of h, then $\omega'(t_1) \geq \omega(t_1)$ $\forall t_1$.)

Proof: Suppose first that $h(t_1) \geq 0$ $\forall t_1$. Let S be the set of all t_1 such that $\omega(t_1) \geq 0$. Then summing (9.2) over all t_1 not in S, we get

$$\sum_{t_1\notin S}\Big(\lambda(t_1) + \sum_{s_1\in S} \alpha_1(s_1|t_1)\Big)\omega(t_1) - \sum_{t_1\notin S}\ \sum_{s_1\in S} \alpha_1(t_1|s_1)\,\omega(s_1) = \sum_{t_1\notin S} h(t_1).$$

The first term here must be strictly negative, unless $S = T_1$. (We use here the fact that every $\lambda(t_1) > 0$, since $\lambda\in\Omega_{++}$.) Since the second term is nonnegative and subtracted, we would have a strictly negative left side equal to a nonnegative right side, unless $S = T_1$. So if all $h(t_1) \geq 0$ then all $\omega(t_1) \geq 0$.

Thus there can be no nonzero solutions to (9.2) if $h = \underset{\sim}{0}$ (since ω and $-\omega$ would both be solutions). So (9.2) is a system of $|T_1|$ independent linear equations in $|T_1|$ unknowns, and it must have a unique solution ω. The fact that ω is componentwise increasing follows from the nonnegativity result of the preceding paragraph, by linearity. Q.E.D.

Lemma 2. For a given incentive problem Γ, suppose that a utility allocation ω is warranted by some λ in Ω_{++} and α in A. Then there exists $\bar{\Gamma}$, an extension of Γ, and there exists μ^*, a safe and undominated mechanism in $\bar{\Gamma}$, such that $\bar{U}_1(\mu^*|t_1) = \omega(t_1)$ for every t_1.

Proof: By Lemma 1, there exists some y in $\mathbb{R}^T$ such that

$$\Big(\lambda(t_1) + \sum_{s_1 \varepsilon T_1} \alpha_1(s_1|t_1)\Big)y(t) - \sum_{s_1 \varepsilon T_1} \alpha_1(t_1|s_1)\, y(t_{-1},s_1)$$

$$= \max_{d \varepsilon D} L(d,t,\lambda,\alpha), \qquad \forall t \varepsilon T.$$

We now construct the extension $\bar{\Gamma}$ by letting

$$\bar{D}_0 = D_0 \cup \{c_0^*\}, \quad \text{and}$$

$$\bar{u}_i(d,t) = \begin{cases} y(t)/p_1(t_{-1}|t_1) & \text{if } i=1 \text{ and } d_0 = c_0^* \\ 0 & \text{if } i \neq 1 \text{ and } d_0 = c_0^* \\ u_i(d,t) & \text{if } d_0 \neq c_0^*. \end{cases}$$

(This definition is the only place where we use the regularity assumption (2.1) in the theory of neutral optima.) Let c^* be any outcome with c_0^* as its enforceable component; and let μ^* be the mechanism such that $\mu^*(c^*|t) = 1$ for every t in T. Then μ^* is safe, because the utility payoffs do not depend on type-reports or private actions, as long as the enforceable action c_0^* is implemented.

Furthermore, μ^* is undominated for the principal in $\bar{\Gamma}$, because it is an optimal solution to the (extended) primal problem for λ. To show this, observe that

$$\sum_{t\in T} \max_{d\in \bar{D}} \bar{L}(d,t,\lambda,\alpha) = \sum_{t\in T} \bar{L}(c^*,t,\lambda,\alpha)$$

$$= \Big(\sum_{t\in T} \big(\lambda(t_1) + \sum_{s_1} \alpha_1(s_1|t_1)\big)\; p_1(t_{-1}|t_1)\; \bar{u}_1(c^*,t)$$

$$- \sum_{t\in T} \alpha_1(t_1|s_1)\; p_1(t_{-1}|s_1)\; \bar{u}_1(c^*,(t_{-1},s_1))\Big)$$

$$= \sum_{t\in T} \lambda(t_1)\; p_1(t_{-1}|t_1)\; \bar{u}_1(c^*,t) = \sum_{t_1\in T_1} \lambda(t_1)\; \bar{U}_1(\mu^*|t_1).$$

(Here $\bar{L}(\cdot)$ is defined by the analogue of (8.3) for $\bar{\Gamma}$ instead of Γ.) Thus, μ^* and α respectively are optimal solutions of the primal and dual problems for λ, in the context of the extended game $\bar{\Gamma}$, because they are feasible for their respective problems and give equal value to the objective functions.

From the definition of $\bar{u}_i(c^*,t)$, it easily follows that

$$\big(\lambda(t_1) + \sum_{s_1\in T_1} \alpha_1(s_1|t_1)\big)\; \bar{U}_1(\mu^*|t_1) - \sum_{s_1\in T_1} \alpha_1(t_1|s_1)\; \bar{U}_1(\mu^*|s_1)$$

$$= \sum_{t_{-1}\in T_{-1}} \max_{d\in D} L(d,t,\lambda,\alpha), \quad \forall t_1\in T_1.$$

So $\bar{U}_1(\mu^*|t_1) = \omega(t_1)$ for every t_1, because ω is the unique allocation vector satisfying the warrant equations (8.10).

Thus μ^* is safe and undominated in $\bar{\Gamma}$ and gives the utility allocation ω to the principal. Q.E.D.

Lemma 3. Suppose that ω is warranted by some λ in Ω_{++} and α in A. Then

$$\omega(t_1) \geq \sum_{t_{-1}\varepsilon T_{-1}} p_1(t_{-1}|t_1) \min_{d\varepsilon D} u_1(d,t), \quad \forall t_1 \varepsilon T_1.$$

Proof. Let $\bar{\Gamma}$, c_0^*, and μ^* be as in the proof of Lemma 2. By Theorem 2, μ^* is an expectational equilibrium in $\bar{\Gamma}$. But if ω violated the inequality in Lemma 3 for some t_1, then this t_1 could do better than μ^* by selecting any mechanism that never used the new enforceable action c_0^*, so μ^* would not be an expectational equilibrium. Q.E.D.

Proof of Theorem 7 (Characterization of neutral optima).

Given the incentive problem Γ, let $C^1(\Gamma)$ be the set of all ω in Ω such that there exist λ in Ω_{++} and α in A by which ω is warranted. Let $C^2(\Gamma)$ be the set of all ω in Ω such that there exists a sequence $\{\omega^k\}_{k=1}^{\infty}$ satisfying $\omega^k \varepsilon C^k$ for each k and

$$\limsup_{k\to\infty} \omega^k(t_1) \leq \omega(t_1), \quad \forall t_1 \varepsilon T_1. \tag{9.3}$$

Let $B^2(\Gamma)$ be the complement of $C^2(\Gamma)$ in Ω; that is $B^2(\Gamma) = \Omega \setminus C^2(\Gamma)$.

By the Strong Solutions and Extensions axiom, together with Lemma 2, no allocation in $C^1(\Gamma)$ can be blocked in Γ. Then by the Openness and Domination axioms, no allocation in $C^2(\Gamma)$ can be blocked in Γ. Thus, $B(\Gamma) \subseteq B^2(\Gamma)$ for any $B(\cdot)$ that satisfies the four axioms.

We now show that $B^2(\cdot)$, as a blocking correspondence, satisfies the four axioms. Domination and Openness are straightforward to check (since $C^2(\Gamma)$ is closed and upward-comprehensive).

To check the Extensions axiom, let $\bar{\Gamma}$ be any extension of Γ. Let ω be any allocation in $C^2(\bar{\Gamma})$, and let $\{\omega^k\}_{k=1}^{\infty}$ be a sequence of allocations in $C^1(\bar{\Gamma})$ that satisfies (9.3). Let λ^k in Ω_{++} and α^k in A be the parameters that warrant ω^k for $\bar{\Gamma}$, and let $\hat{\omega}^k$ be the allocation warranted by λ^k and α^k for Γ. Then for every t_1,

$$\left(\lambda^k(t_1) + \sum_{s_1} \alpha_1^k(s_1|t_1)\right)\omega^k(t_1) - \sum_{s_1} \alpha_1^k(t_1|s_1)\,\omega^k(s_1)$$

$$= \sum_{t_{-1}} \max_{d \in \bar{D}} \bar{L}(d,t,\lambda,\alpha)$$

$$\geq \sum_{t_{-1}} \max_{d \in D} L(d,t,\lambda,\alpha)$$

$$= \left(\lambda^k(t_1) + \sum_{s_1} \alpha_1^k(s_1|t_1)\right)\hat{\omega}^k(t_1) - \sum_{s_1} \alpha_1(t_1|s_1)\,\hat{\omega}^k(t_1),$$

since $\bar{L}$ is the extension of L to the larger domain $\bar{D} \supseteq D$. Thus, by Lemma 1, $\omega^k(t_1) \geq \hat{\omega}^k(t_1)$ for every t_1, and so $\{\hat{\omega}^k\}_{k=1}^{\infty}$ is a sequence of allocations in $C^1(\Gamma)$ that satisfies (9.3) for ω. Thus $\omega \in C^2(\Gamma)$. So $C^2(\Gamma) \supseteq C^2(\bar{\Gamma})$, and $B^2(\Gamma) \subseteq B^2(\bar{\Gamma})$.

To check the Strong Solutions axiom, suppose that μ is safe and undominated in Γ. There exists some λ in Ω_{++} such that μ is an optimal solution of the primal problem for λ. Let α be an optimal solution of the dual problem for λ, and let ω be the principal's allocation warranted by λ and α. By Lemma 2, there is an extension of Γ in which some safe and undominated mechanism μ^* gives the principal the expected utility allocation ω. But μ (extended by giving zero probability to the new outcomes in $\bar{D} \setminus D$) is still a safe mechanism in the extension of Γ. So $\omega(t_1) \geq U_1(\mu|t_1)$ for all t_1, because Theorem 1 would require that μ could not be incentive compatible given the types in T_1 that prefer μ strictly over μ^*, if any such types existed. This implies that $U_1(\mu) \in C^2(\Gamma)$, so $U_1(\mu) \notin B^2(\Gamma)$.

Thus $B^2(\cdot)$ is the maximal blocking correspondence that satisfies the four axioms, and so $B^2(\Gamma) = B^*(\Gamma)$, in the notation of Section 7. Thus μ is a neutral optimum for the principal in Γ if and only if $U_1(\mu)\notin B^2(\Gamma)$, or equivalently, if and only if $U_1(\mu)\in C^2(\Gamma)$. The conditions in Theorem 7 restate the definition of $C^2(\Gamma)$. Q.E.D.

Proof of Theorem 8 (Necessary conditions for a neutral optimum). Let $\{\lambda^k,\alpha^k,\omega^k\}_{k=1}^{\infty}$ satisfy the conditions of Theorem 7 for the neutral optimum μ. Since the warrant equations (8.8) are linearly homogeneous in λ^k and α^k, we may assume without loss of generality that each (λ^k,α^k) pair is in some closed and bounded set that excludes $(\underset{\sim}{0},\underset{\sim}{0})$ in $\Omega_+ \times A$. (For example, we could require that $\|\lambda^k\| + \|\alpha^k\| = 1 \quad \forall k$.) Choosing a subsequence if necessary, we can also assume that the $\{\lambda^k\}$ and $\{\alpha^k\}$ sequences are convergent to some (λ,α) such that $(\lambda,\alpha) \neq (\underset{\sim}{0},\underset{\sim}{0})$. By Lemma 3 and (8.9), the $\{\omega^k\}$ sequence is also bounded, so we can also assume that it is convergent to some limit ω. By summing (8.8) over all t_1, wet get

$$\sum_{t_1\varepsilon T_1} \lambda^k(t_1)\, \omega^k(t_1) = \sum_{t\varepsilon T} \max_{d\varepsilon D} L(d,t,\lambda^k,\alpha^k), \quad \forall k.$$

Then taking limits as $k\to\infty$ and applying (8.9), we get $\omega(t_1) \leq U_1(\mu|t_1)\ \forall t_1$, and

$$(9.4) \qquad \sum_{t_1\varepsilon T_1} \lambda(t_1)\, U_1(\mu|t_1) \geq \sum_{t_1\varepsilon T_1} \lambda(t_1)\, \omega(t_1) = \sum_{t\varepsilon T} \max_{d\varepsilon D} L(d,t,\lambda,\alpha).$$

But μ is feasible in the primal for λ, and α is feasible in the dual for λ. So by duality theory, μ and α are optimal solutions of the primal and dual problems for λ, respectively, and the inequality (9.4) must be an equality,

$$(9.5) \qquad \sum_{t_1\varepsilon T_1} \lambda(t_1)\, \omega(t_1) = \sum_{t_1\varepsilon T_1} \lambda(t_1)\, U_1(\mu|t_1),$$

Equation (9.5) implies the complementary slackness conditions in (8.14), since each $\lambda(t_1) \geq 0$). The limit of (8.8) gives us (8.13). Q.E.D.

Proof of Theorem 6 (Existence of neutral optima). To prove the existence of neutral optima, we begin with some definitions. Let Λ be the unit simplex in $\Omega = \mathbb{R}^{T_1}$,

$$\Lambda = \{\lambda \varepsilon \Omega_+ | \sum_{t_1 \varepsilon T_1} \lambda(t_1) = 1\}.$$

For any k larger than $|T_1|$, let

$$\Lambda^k = \{\lambda \varepsilon \Lambda | \lambda(t_1) \geq 1/k, \quad \forall t_1 \varepsilon T_1\}.$$

We let F denote the set of all incentive-compatible mechanisms for Γ.

There exists a compact convex set A^* such that $A^* \subseteq A$ and, for each λ in Λ, A^* contains at least one optimal solution of the dual problem for λ. To prove this fact, observe that F, the feasible set of the primal problem, is compact and independent of λ. So the simplex Λ can be covered by a finite collection of sets (each set corresponding to the range of optimality of one basic feasible solution in the primal) such that, within each set, an optimal solution of the dual can be given as a linear function of λ. Each of these linear functions is bounded on Λ, so we can let A^* be the convex hull of the union of the ranges of these linear functions on Λ.

For any k greater than $|T_1|$, we now define a correspondence $Z^k: F \times A^* \times \Lambda^k \Rightarrow F \times A^* \times \Lambda^k$ so that $(\mu'', \alpha'', \lambda'') \varepsilon Z^k(\mu', \alpha', \lambda')$ iff

μ'' is an optimal solution of the primal for λ';

α'' is an optimal solution of the dual for λ'; and

$\lambda'' = 1/k$ for each t_1 such that

$$\omega'(t_1) - U_1(\mu'|t_1) < \max_{s_1 \varepsilon T_1} (\omega'(s_1) - U_1(\mu'|s_1)),$$

where ω' is the allocation warranted by λ' and α'.

That is, λ'' must put as much weight as possible on the types whose claims warranted by λ' and α' most exceed their actual allocation from μ'.

By the Kakutani fixed point theorem, for each k there exists some $(\mu^k, \alpha^k, \lambda^k)$ such that

$$(\mu^k, \alpha^k, \lambda^k) \varepsilon Z^k(\mu^k, \alpha^k, \lambda^k).$$

Since this sequence of fixed points is in a compact domain, we can choose a convergent subsequence, converging to some (μ, α, λ) in $F \times A^* \times \Lambda$. We now show that this μ is a neutral optimum for the principal in the incentive problem Γ.

Let ω^k be the principal's allocation that is warranted by λ^k and α^k. By the warrant equations and duality theory (as in (9.5)),

$$\sum_{t_1 \varepsilon T_1} \lambda^k(t_1)\, \omega^k(t_1) = \sum_{t_1 \varepsilon T_1} \lambda^k(t_1)\, U_1(\mu^k|t_1).$$

For any t_1, if $\omega^k(t_1) < U_1(\mu^k|t_1)$ then $\lambda^k(t_1) = 1/k$. So for any t_1,

$$\text{if} \quad \liminf_{k\to\infty} \omega^k(t_1) < U_1(\mu|t_1) \quad \text{then} \quad \lim_{k\to\infty} \lambda^k(t_1) = 0.$$

Now suppose that there were some s_1 in T_1 such that

$$\limsup_{k\to\infty} \omega^k(t_1) > U_1(\mu|s_1) = \lim_{k\to\infty} U_1(\mu^k|s_1).$$

Then we could find some such s_1 for which $\lambda(s_1) > 0$ also. But then

$$0 < \limsup_{k\to\infty} \lambda^k(s_1)\big(\omega^k(s_1) - U_1(\mu^k|s_1)\big)$$

$$= \limsup_{k\to\infty} \sum_{t_1 \neq s_1} \lambda^k(t_1)\big(U_1(\mu^k|t_1) - \omega^k(t_1)\big) \leqslant 0,$$

which is impossible. So no such s_1 can exist, and so for every t_1

$$\limsup_{k\to\infty} \omega^k(t_1) \leq U_1(\mu|t_1).$$

Thus $\{\lambda^k,\alpha^k,\omega^k\}_{k=1}^{\infty}$ satisfy the conditions of Theorem 7 for μ. Q.E.D.

Proof of Theorem 5. We must show that expectational equilibria and core mechanisms can both be characterized as the set of incentive-compatible mechanisms that give "unblocked" allocations, in terms of some blocking concepts that satisfy the four axioms.

Given an incentive problem Γ, we define $B^C(\Gamma)$ so that $\omega \varepsilon B^C(\Gamma)$ iff there exists some mechanism ν and some nonempty set R such that $R \subseteq T_1$, $\omega(t_1) < U_1(\nu|t_1)$ for every t_1 in R, and ν is incentive compatible given S, for every S such that $R \subseteq S \subseteq T_1$. Thus μ is a core mechanism if and only if μ is incentive compatible and $U_1(\mu) \notin B^C(\Gamma)$. It is straightforward to check that $B^C(\cdot)$ satisfies the Domination and Openness axioms. The Extensions axiom holds, because any mechanism that is incentive compatible given S in Γ is also incentive compatible given S in any extension of Γ. By Theorem 1, strong solutions are core mechanisms, so $B^C(\cdot)$ satisfies the Strong Solutions axiom. So $B^C(\Gamma) \subseteq B^*(\Gamma)$, and every neutral optimum is a core mechanism.

We define $B^E(\Gamma)$ so that $\omega \varepsilon B^E(\Gamma)$ iff there exists some mechanism ν such that, for every (γ,τ,Q), if (γ,τ) is a Nash equilibrium of ν given Q then there exists some t_1 in T_1 such that $\omega(t_1) < W_1(\nu,\gamma,\tau|t_1,Q)$. Thus, μ is an expectational equilibrium if and only if $U_1(\mu) \notin B^E(\Gamma)$. By Theorem 2, $B^E(\cdot)$ satisfies the Strong Solutions axiom. The Extensions axiom holds because if (γ,τ) is a Nash equilibrium for ν given Q in Γ then the same is true in any extension of Γ (since the extension differs from Γ only by the addition of new enforceable actions which ν does not use). The Domination axiom is obvious for $B^E(\cdot)$.

To prove that $B^E(\Gamma)$ is open, we show that the complement is closed. Suppose that $\{\omega^k\}_{k=1}^{\infty}$ is a sequence of allocations that are not in $B^E(\Gamma)$ and that converge to some ω. Given any ν, for every k there exists some normalized-likelihood vector Q^k and some Nash equilibrium (γ^k,τ^k) for ν given Q^k, such that $\omega^k(t_1) \geqslant W_1(\nu,\gamma^k,\tau^k|t_1,Q^k)$ for every t_1. Choosing a subsequence if necessary, the (γ^k,τ^k,Q^k) converge to some (γ,τ,Q) such that (γ,τ) is a Nash equilibrium for ν given Q and $\omega(t_1) \geqslant W_1(\nu,\gamma,\tau|t_1,Q)$ for every t_1. This construction is possible for every ν, so $\omega \notin B^E(\Gamma)$. Thus $B^E(\cdot)$ satisfies all four axioms, and any neutral optimum is an expectational equilibrium. Q.E.D.

<u>Proof of Theorems 3 and 4</u>. By Theorem 6, a neutral optimum exists. By Theorem 5, a neutral optimum is a core mechanism and an expectational equilibrium. So there exists a core mechanism and an expectational equilibrium.

REFERENCES

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

Bhattacharya, S., "Signaling Environments: Efficiency,Sequential Equilibria, and Myerson's Conjecture," mimeo, Stanford University (1981).

Dasgupta, P. C., P. J. Hammond, and E. S. Maskin, "The Implementation of Social Choice Rules: Some Results on Incentive Compatibility," Review of Economic Studies 46 (1979), 185-216.

Gibbard, A., "Manipulation of Voting Schemes: a General Result," Econometrica 41 (1973), 587-602.

Harris, M., and A. Raviv, "Optimal Incentive Contracts with Imperfect Information," Journal of Economic Theory 20 (1979), 231-259.

Harris, M., and R. M. Townsend, "Resource Allocation Under Asymmetric Information," Econometrica 49 (1981), 33-64.

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

Holmström, B., "On Incentives and Control in Organizations" Ph.d. Dissertation, Stanford University (1977).

Holmström, B., "Moral Hazard and Observability," Bell Journal of Economics 10 (1979), 74-91.

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

Mirrlees, J., "The Optimal Structure of Incentives and Authority within an Organization," Bell Journal of Economics 7 (1976), 165-181.

Myerson, R. B., "Incentive Compatibility and the Bargaining Problem," Econometrica 47 (1979), 61-73.

Myerson, R. B., "Optimal Coordination Mechanisms in Generalized Principal-Agent Problems," Journal of Mathematical Economics 10 (1982a), 67-81.

Myerson, R. B., "Two-Person Bargaining Problems with Incomplete Information," mimeo, Northwestern University (1982b).

Myerson, R. B., "Cooperative Games with Incomplete Information," mimeo, Northwestern University (1982c).

Riley, J. G.,"Informational Equilibrium," Econometrica 47 (1979), 331-359.

Rosenthal, R. W., "Arbitration of Two-Party Disputes under Uncertainty," Review of Economic Studies 45 (1978), 595-604.

Ross, S., "The Economic Theory of Agency: the Principal's Problem," American Economic Review 63 (1973), 134-139.

Roth, A. E., Axiomatic Models of Bargaining, Berlin: Springer-Verlag (1979).

Rothschild, M., and J. Stiglitz, "Equilibrium in Competitive Insurance Markets: An Essay on the Economics of Imperfect Information," Quarterly Journal of Economics 90 (1976), 629-649.

Wilson, C., "A Model of Insurance Markets with Incomplete Information," Journal of Economic Theory 16 (1977), 167-207.

Wilson, C., "The Nature of Equilibrium in Markets with Adverse Selection," Bell Journal of Economics 11 (1980), 108-130.