Journal of Mathematical Economics 10 (1982) 67–81. North-Holland Publishing Company

OPTIMAL COORDINATION MECHANISMS IN GENERALIZED PRINCIPAL–AGENT PROBLEMS

Roger B. MYERSON*

Northwestern University, Evanston, IL 60201, USA

The general principal–agent problem is formulated, in which agents have both private information and private decisions, unobservable to the principal. It is shown that the principal can restrict himself to incentive-compatible direct coordination mechanisms, in which agents report their information to the principal, who then recommends to them decisions forming a correlated equilibrium. In the finite case, optimal coordination mechanisms can be found by linear programming. Some basic issues relating to systems with many principals are also discussed. Non-cooperative equilibria between interacting principals do not necessarily exist, so quasi-equilibria are defined and shown to exist.

1. Introduction

This is a paper about game design. The central question to be considered is how an individual should structure a social situation which he controls, so as to maximize his expected utility. Following the terminology of Ross (1973), we refer to the individual in control as the *principal*, and we refer to the other individuals as the *agents*. That is, we think of an individual as a principal if he is in a position to set the rules of communication and he can establish the structure of incentives to which the other individuals must react. In short, the principal is the individual who has the power to design the game which the other individuals (or agents) must play.

The use of the terms 'principal/agent' here may require some explanation. Possible alternative terms could have been 'leader/follower', or 'exploiter/exploited', or 'sovereign/subject'. I have chosen to use the principal/agent terminology because it suggests a class of important applications in which economists consider it acceptable for one individual (such as the owner of a corporation) to design a social coordination system (the corporation) so as to optimally exploit some other individuals (his managers and employees). However, we will here characterize the individuals' decisions and information at a level of abstraction in which we lose much of the structure used in most of the literature on principal–agent problems. See

*The author is indebted to Rick Antle, Paul Milgrom, and especially to Bengt Holmström for valuable discussion and suggestions.

0304–4068/82/0000–0000/$02.75

Mirrlees (1976), Harris and Raviv (1979), and Holmström (1979, 1980) for some of the important results and applications of principal–agent analysis; Holmström (1980) offers an excellent survey. The problems to be considered in this paper should be properly called *generalized* principal–agent problems, to distinguish them from the more specific principal–agent structures studied in this literature.

Our generalized principal–agent problem may be viewed as a hybrid between a cooperative and a non-cooperative game. The agents are assumed to act non-cooperatively as utility maximizers who will passively accept any Nash equilibrium of any game which the principal chooses. The possibilities for communication and cooperation are assumed to be entirely controlled by the principal. Thus, when we identify a particular individual as the 'principal', we mean that we want to know what would be the best possible cooperative arrangement for the purposes of this individual, if he had all of the bargaining ability.

Our generalized principal–agent problem can also be interpreted as a social choice problem, where the principal is a social planner and his utility function is a social welfare function. Then the principal's mechanism-design problem can be reinterpreted as the problem of finding a coordination mechanism for the agents which maximizes the expected social welfare.

In designing a game or coordination mechanism, a principal faces two constraining factors: the agents may have private information which he cannot directly observe; and the agents may have private decision domains which he cannot directly control. In the insurance literature, these two factors are known as *adverse selection* and *moral hazard*. In order to get information from the agents and influence their decisions, the principal must design a coordination system which gives the agents the incentive to do as he intends. That is, his mechanism must be *incentive compatible*.

In section 2, we develop the formal structure of our generalized principal–agent problem. In this formulation, we follow the Bayesian approach to the study of games with incomplete information, as suggested by Harsanyi (1967/8). We use the Bayesian equilibrium as the solution concept for such games. To compute the equilibria of any game can often be extremely difficult, so it may be remarkable that the class of *all* outcomes which the principal can achieve as Bayesian equilibria of games is actually quite easy to characterize. We show this result in section 3. The key insight is that the principal may, without loss of generality, restrict himself to incentive-compatible direct coordination mechanisms.

This basic insight has been independently discovered by many authors for the case where agents have private information; see Gibbard (1973), Rosenthal (1978), Holmström (1977), Dasgupta, Hammond and Maskin (1979), Myerson (1979), and Harris and Townsend (1981). In Myerson (1981), this insight was referred to as the *revelation principle*, since it asserts that

there is no loss of generality in assuming that the principal should structure his incentive system so that all agents will be willing to reveal all of their information to him honestly. The key contribution of Proposition 2 in section 3 is to extend the revelation principle to situations where there are moral hazard factors, in addition to informational or adverse selection factors, in the principal's problem. Proposition 2 can be best viewed as a synthesis of the model of Holmström (1977) (where moral hazard effects were also considered) with Aumann's (1974) concept of *correlated equilibrium.* General principal–agent problems with both moral hazard and adverse selection effects have been studied, in the context of auditor contracts, by Antle (1980).

In section 4, we consider situations in which several principals interact in designing coordination mechanisms for their agents. The obvious notion of equilibrium among principals unfortunately does not have the existence property, so we define a modified concept of quasi-equilibrium among principals, which can be shown to always exist.

2. General formulation of the principal–agent problem

In the generalized principal–agent problem, there is one principal and there are n agents, numbered 1 to n. There are two reasons why the principal may need to rely on these agents: they may have information which he cannot observe, or they may have the power to make some decisions which he cannot directly control. For each agent i, we let T_i denote the set of all possible states of agent i's private information, and we let D_i denote the set of all possible private decisions or actions which agent i can make. Following Harsanyi (1967/8), we refer to T_i as the set of all possible *types* for agent i, where each type t_i in T_i represents a complete description of all the private information i might have about his environment, his abilities, and his preferences. Each private decision-option d_i in D_i may represent, for example, a level of effort which agent i might exert in working for the principal, and which the principal cnnot observe or control.

We let D_0 denote the principal's decision domain. Each option d_0 in D_0 may represent a description of how the principal might plan to allocate his resources to the productive agents, or how he might plan to reward them as a function of some future observations about output, etc. Any actions by an agent which the principal can observe and control directly should also be considered as part of the principal's decision domain.

We shall use the notation

$$D = D_0 \times D_1 \times \cdots \times D_n, \qquad T = T_1 \times \cdots \times T_n.$$

That is, D is the set of all combinations of decisions which could be made by

the principal and the agents, and T is the set of all possible combinations of agents' types. We let $U_0: D \times T \to \boldsymbol{R}$ denote the principal's utility function, and we let $U_i: D \times T \to \boldsymbol{R}$ denote the utility function of agent i. That is, for any $(d,t)=(d_0,d_1,\ldots,d_n,t_1,\ldots,t_n)$ in $D \times T$, $U_i(d,t)$ represents the expected utility payoff for agent i, measured in some von Neumann–Morgenstern utility scale, if the principal and the agents act according to the vector of decisions d, and if the agents' information is as represented by the vector of types t. Any random variable not observed by any agents should be integrated out, using their conditional distributions given t, to compute this expected utility $U_i(d,t)$.

We let P denote the probability distribution on T, so that $P(t)$ is the probability (as would be assessed by the principal, or by any agent *ex ante*) that $t=(t_1,\ldots,t_n)$ will be the vector of agents' types. For mathematical simplicity, we shall generally assume that D_0 and all D_i and T_i sets are finite.

Given these structures $(D_0, U_0, (D_i, T_i, U_i)_{i=1}^n, P)$, the principal's problem is to coordinate his decisions and those of his agents, so as to maximize his expected utility. We assume that the principal has complete control over all communication between the agents, that he can request any information which the agents are willing to send, and that he can send messages and recommendations to the agents. However, the principal cannot directly observe an agent's type in T_i or control the decisions in any D_i, except D_0.

This formulation is rather abstract, so it may be worth seeing how it applies to the conventional principal–agent problem. For example, suppose that there is one agent who gets some private information and then chooses an effort level which affects the quantity of output accruing to the principal. Then $n=1$, and the agent's private decision d_1 is his effort level. If the principal can observe the quantity of output before paying the agent, then the principal's decision d_0 is itself a function mapping observed outputs to salary levels for the agent, and so D_0 is a set of functions from $\boldsymbol{R}$ to $\boldsymbol{R}$. If the quantity of output is $f(d_1,\theta)$, depending on the agent's effort d_1 and on some random variable θ which is correlated with the agent's information t_1, then the agent's utility function in our formulation is

$$U_1(d_0,d_1,t_1)=E[u_1(d_0(f(d_1,\theta)),d_1)\,|\,t_1],$$

where $u_1(\cdot)$ is the agent's utility for money and effort, and the expectation is taken over θ given t_1. Similarly, if $u_0(\cdot)$ is the principal's utility for money, then

$$U_0(d_0,d_1,t_1)=E[u_0(f(d_1,\theta)-d_0(f(d_1,\theta)))\,|\,t_1].$$

To describe a typical coordination mechanism which could be established by the principal, we let M_i be the set of all possible messages which agent i

might receive from the principal or from the other agents in the mechanism. We let R_i be the set of all strategies which agent i could use for sending reports to the principal or the other agents in the mechanism, given his own type. (We are not assuming that this communication system is necessarily a one-stage affair. To model a multi-stage communication system, let R_i be the set of all possible strategic plans which agent i might use to determine the reports he sends at each stage as a function of the messages he has received at earlier stages, and let M_i be the set of all possible sequences of messages which agent i might receive over the whole process.) Of course, the messages received by one agent must depend on the reports sent by the others, and the principal's decision in D_0 may also depend on these reports. To describe this dependence, let

$$\pi(d_0, m_1, \ldots, m_n \mid r_1, \ldots, r_n)$$

be the probability that the principal will choose d_0 in D_0 and that each agent i will receive message m_i in M_i, given that each agent i is planning to send reports according to the reporting strategy r_i in R_i. Then, these sets $(R_i, M_i)_{i=1}^n$ together with the probability function π completely describe the coordination mechanism established by the principal.

In the context of this coordination mechanism $((R_i, M_i)_{i=1}^n, \pi)$ each agent i controls his choice of reporting strategy in R_i as a function of his type, and controls his choice of a decision in D_i as a function of his type and his message received. That is, agent i selects a pair of functions $\rho_i: T_i \to R_i$ and $\delta_i: M_i \times T_i \to D_i$, such that $\rho_i(t_i)$ would be i's reporting strategy if i were of type t_i, and $\delta_i(m_i, t_i)$ would be i's final decision in D_i after he received message m_i if his type were t_i. We may refer to any pair of such functions (ρ_i, δ_i) as a *participation strategy* for i.

Given any n-tuple of participation strategies $((\rho_1, \delta_1), \ldots, (\rho_n, \delta_n))$ for the n agents, the expected utility for agent i would be

$$V_i((\rho_1, \delta_1), \ldots, (\rho_n, \delta_n)) = \sum_{t \in T} \sum_{d_0 \in D_0} \sum_{m \in M} P(t)\, \pi(d_0, m \mid \rho(t))\, U_i(d_0, \delta(m, t), t), \tag{1}$$

where

$$M = M_1 \times \cdots \times M_n,$$

$$\rho(t) = (\rho_1(t_1), \ldots, \rho_n(t_n)),$$

$$(\delta(m, t)) = (\delta_1(m_1, t_1), \ldots, \delta_n(m_n, t_n)).$$

Thus, the principal's coordination mechanism, in effect, defines a game to be played by the agents. Since each agent's participation strategy is chosen

independently (and in particular, cannot be dictated by the principal), the principal should expect the agents to use participation strategies which form a non-cooperative equilibrium to this game, in which every agent's participation strategy maximizes his expected utility. Formally, the participation strategies $((\rho_1, \delta_1), \ldots, (\rho_n, \delta_n))$ form an *equilibrium* if and only if, for every agent i and every alternative participation strategy $(\tilde{\rho}_i, \tilde{\delta}_i)$ for i,

$$V_i((\rho_1, \delta_1), \ldots, (\rho_i, \delta_i), \ldots, (\rho_n, \delta_n)) \geqq V_i(\rho_1, \delta_1), \ldots, (\tilde{\rho}_i, \tilde{\delta}_i), \ldots, (\rho_n, \delta_n)). \quad (2)$$

The principal's problem is to design a coordination mechanism $((R_i, M_i)_{i=1}^n, \pi)$ such that there is an equilibrium of participation strategies $((\rho_i, \delta_1), \ldots, (\rho_n, \delta_n))$ which gives the principal the highest possible expected utility $V_0((\rho_1, \delta_1), \ldots, (\rho_n, \delta_n))$, where V_0 is defined as in (1) for $i=0$.

3. Direct coordination mechanisms and incentive compatibility

The principal's problem might seem to be intractable, because we have put no bound on the complexity of the sets M_i and R_i. However, we can now show that there is no loss of generality in restricting our attention to communication and control mechanisms with a very simple structure.

Following the terminology of Dasgupta, Hammond and Maskin (1979), we say that a coordination mechanism is *direct* iff each $M_i = D_i$ and each $R_i = T_i$. That is, in a direct coordination mechanism, each agent i is asked to report a type in T_i to the principal, and in return, the principal will send each agent i a suggested decision in D_i. Thus, a direct mechansim is completely characterized by its probability function π, where $\pi(d_0, d_1, \ldots, d_n \mid t_1, \ldots t_n)$ is the principal's probability of doing d_0 and recommending d_i to each agent i if each agent i reports t_i.

In a direct mechanism, agent i is *honest and obedient* if he uses the participation strategy (ρ_i^*, δ_i^*) satisfying, for all t_i and d_i,

$$\rho_i^*(t_i) = t_i \quad \text{and} \quad \delta_i^*(d_i, t_i) = d_i.$$

We say that a direct mechanism π is *incentive-compatible* [or, more correctly, *Bayesian incentive-compatible*, in the sense of D'Aspremont and Gerard-Varet (1979)] iff the honest-obedient participation strategies $((\rho_1^*, \delta_1^*), \ldots, (\rho_n^*, \delta_n^*))$ form an equilibrium.

More precisely, we may characterize the set of incentive compatible direct mechanisms as the set of all functions $\pi: D \times T \to \boldsymbol{R}$ satisfying the following inequalities:

$$\pi(d \mid t) \geqq 0 \quad \text{and} \quad \sum_{e \in D} \pi(e \mid t) = 1, \qquad \forall\, d \in D, \quad \forall\, t \in T, \quad (3)$$

and

$$\sum_{\substack{t\in T\\ t_i=\tau_i}}\sum_{d\in D} P(t)\,\pi(d\,|\,t)\,U_i(d,t)$$
$$\geqq \sum_{\substack{t\in T\\ t_i=\tau_i}}\sum_{d\in D} P(t)\,\pi(d\,|\,t_{-i},\hat{\tau}_i)\,U_i((d_{-i},\hat{\delta}_i(d_i)),t),$$
$$\forall\, i\in\{1,\ldots,n\},\quad \forall\,\tau_i\in T_i,\quad \forall\,\hat{\tau}_i\in T_i,\quad \forall\,\hat{\delta}_i: D_i\to D_i. \tag{4}$$

[We use here the notation $(t_{-i},\tau_i)=(t_1,\ldots,\tau_i,\ldots,t_n)$ and $(d_{-i},\hat{\delta}_i(d_i))=(d_0,d_1,\ldots,\hat{\delta}_i(d_i),\ldots,d_n)$. The first summation in (4) is over all t in T such that $t_i=\tau_i$.] Condition (3) asserts that π is a valid conditional probability function. Condition (4) asserts that when agent i is type τ_i, he should do at least as well by being honest and obedient as by reporting $\hat{\tau}_i$ and then using $\hat{\delta}_i(d_i)$ when told to do d_i, given that all other agents are planning to be honest and obedient.

When the T_i and D_i sets are finite, the set of incentive-compatible direct coordination mechanisms is characterized by finitely many linear inequalities, since (3) and (4) are linear in π, and the number of mappings $\delta_i: D_i\to D_i$ is finite. The principal's expected utility from the direct mechanism π is

$$\sum_{t\in T}\sum_{d\in D} P(t)\,\pi(d\,|\,t)\,U_0(d,t), \tag{5}$$

if all agents are honest and obedient. Formula (5) is linear in π. Thus, we are led to the following important observation.

Proposition 1. When the type sets T_i and the decision sets D_i (including D_0) are all finite, then the problem of computing the optimal incentive compatible direct coordination mechanism is a linear programming problem: to choose $\pi: D\times T\to \boldsymbol{R}$ so as to maximize (5) *subject to* (3) *and* (4).

We may now ask whether other coordination mechanisms could possibly offer the principal higher expected utility than the best incentive-compatible direct mechanism. Perhaps an equilibrium of participation strategies in which some lying or disobedience is accepted might be better, or some coordination mechanism with $R_i\neq T_i$ or $M_i\neq D_i$ might be better. In fact, the answer is No.

Proposition 2. Given any equilibrium of participation strategies $(\rho_i,\delta_i)_{i=1}^n$ in any coordination mechanism $((R_i,M_i)_{i=1}^n,\pi)$, there exists an incentive-compatible direct mechanism π^* in which the principal gets the same expected utility (when the agents are honest and obedient) as in the given equilibrium of the given mechanism. Thus, the optimal incentive-compatible direct coordination mechanism is also optimal in the class of all coordination mechanisms.

Proof. Given the equilibrium of participation strategies $(\rho_i, \delta_i)_{i=1}^n$, let $\delta^{-1}(d,t)$ be the set of all messages to the agents such that each agent i would respond by choosing decision d_i if his type were t_i. That is,

$$\delta^{-1}(d,t) = \{m \mid \delta_i(m_i, t_i) = d_i, \text{ for all } i\}.$$

Then, define $\pi^*: D \times T \to \boldsymbol{R}$ so that

$$\pi^*(d \mid t) = \sum_{m \in \delta^{-1}(d,t)} \pi(d_0, m \mid \rho_1(t_1), \ldots, \rho_n(t_n)).$$

π^* is the direct coordination mechanism which simulates the overall effect of the original mechanism with the given participation strategies, as illustrated in fig. 1. That is, $\pi^*(d \mid t)$ is the probability that the principal will choose d_0 and each agent will choose d_i if t is the vector of types, when each agent i chooses his reporting strategy according to ρ_i, the principal's decision and the messages to the agents are determined from these reports by π, and each agent i uses δ_i to translate his received message into a decision in D_i.

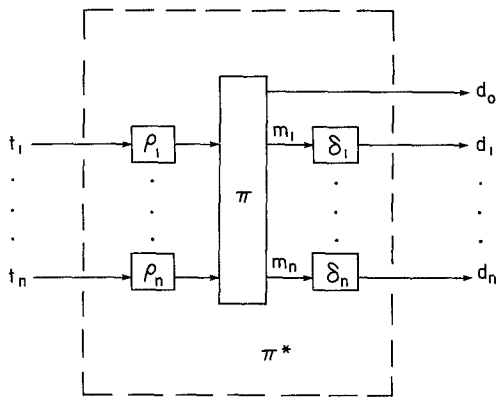

Fig. 1

It is easy to see that π^* gives the same expected utility to the principal (and to each agent) as the originally given mechanism, since the probability distribution over decision vectors for any type vector is the same. To check that π^* is incentive-compatible, suppose (to the contrary) that agent i could gain when τ_i is his type by reporting $\hat{\tau}_i$ and then by permuting his instructions by $\hat{\delta}: D_i \to D_i$, in violation of (4). Then, consider the participation strategies $(\tilde{\rho}_i, \tilde{\delta}_i)$ for the original mechanism, where

$$\tilde{\rho}_i(\tau_i) = \rho_i(\hat{\tau}_i), \qquad \tilde{\rho}_i(t_i) = \rho_i(t_i) \qquad \text{if} \quad t_i \neq \tau_i,$$

$$\tilde{\delta}_i(m_i, \tau_i) = \hat{\delta}(\delta_i(m_i, \hat{\tau}_i)), \quad \tilde{\delta}_i(m_i, t_i) = \delta_i(m_i, t_i) \quad \text{if} \quad t_i \neq \tau_i.$$

That is, $(\tilde{\rho}_i, \tilde{\delta}_i)$ differs from (ρ_i, δ_i) in that, when τ_i is i's true type, he sends reports as if $\hat{\tau}_i$ were his type and then he permutes his planned responses by δ from what he would have done if $\hat{\tau}_i$ were his type. It is straightforward to check that if $(\hat{\tau}_i, \hat{\delta})$ violated (4) for π^*, then $(\tilde{\rho}_i, \tilde{\delta}_i)$ would violate (2) for the originally given mechanism, which would contradict the fact that (ρ_i, δ_i) is an equilibrium strategy for i. Thus, π^* must be incentive compatible. Q.E.D.

Consider the special case where $|D_0| = 1$ and each $|T_i| = 1$, so that the principal does not control any decisions directly (he has only one feasible decision) and each agent i has no private information (he could only be of one type). In this case, the incentive-compatible mechanisms are *correlated equilibria*, in the sense of Aumann (1974). That is, the agents are playing a non-cooperative game, and the principal, by his control over all communication between agents, can select the correlated equilibrium which maximizes his own expected utility.

It is worth noting that Proposition 2 does not depend in any way on the finiteness of T_i and D_i. Even if the T_i, D_i, R_i, and M_i sets were infinite sets, we could still construct an incentive-compatible direct coordination mechanism to simulate any equilibrium of measurable participation strategies. The essential ideas in fig. 1 carry over to the infinite case, except that the simple probability vectors must be replaced by measures. That is $\pi(\cdot \mid \rho_1(t_1), \ldots, \rho_n(t_n))$ is now a probability measure on $D_0 \times M_1 \times \cdots \times M_n$, and $\pi^*(\cdot \mid t)$ must be a measure on D defined so that, for any measurable set $C \subseteq D$,

$$\pi^*(C \mid t) = \pi(\delta^{-1}(C, t) \mid \rho_1(t_1), \ldots, \rho_n(t_n)),$$

where

$$\delta^{-1}(C, t) = \{(d_0, m_1, \ldots, m_n) \mid (d_0, \delta_1(m_1, t_1), \ldots, \delta_n(m_n, t_n)) \in C\}.$$

We have emphasized the finite case in this section mainly to simplify the logical exposition, and because it is only in the finite case that the principal's problem reduces to a linear programming problem.

4. Equilibria among several principals

A difficulty arises when we try to construct a general theory of principal–agent interaction among several principals, each with his own set of agents. Suppose that there are N principals, numbered 1 to N, with k denoting a typical principal. Each principal k has n_k agents, numbered 1 to n_k, with i denoting a typical agent. We shall assume that the sets of agents for different principals are disjoint, so that each agent belongs to just one principal. Principal k together with his agents may be referred to as *corporation k*. Extending the notation of section 2, we let T_i^k and D_i^k denote the sets of types

and decisions for the ith agent belonging to principal k. We let D_0^k denote the decision domain directly controlled by principal k. We use the notation:

$$D^k = D_0^k \times D_1^k \times \cdots \times D_{n_k}^k, \qquad T^k = T_1^k \times \cdots \times T_{n_k}^k,$$

$$D = D^1 \times \cdots \times D^N, \qquad T = T^1 \times \cdots \times T^N.$$

The utility functions of the kth principal and his ith agent are $U_0^k: D \times T \to \boldsymbol{R}$ and $U_i^k: D \times T \to \boldsymbol{R}$ respectively. Thus, we are allowing that the utility payoff to any principal or any agent may depend on the decisions and types in all the corporations. For any t in T, we let $P(t)$ denote the probability that t is the true vector of types for all the agents.

As in section 3, we may define a *direct mechanism* for corporation k to be any function $\pi^k: D^k \times T^k \to \boldsymbol{R}$ where $\pi^k(d^k \mid t^k)$ is the probability that the principal will use decision d_0^κ and recommend d_i^k to his ith agent, for each i, if t^k is the vector of reports from his n_k agents. [Here, $t^k = (t_1^k, \ldots, t_{n_k}^k)$ and $d^k = (d_0^k, d_1^k, \ldots, d_{n_k}^k)$.] Thus, π^k must satisfy

$$\pi^k(d^k \mid t^k) \geqq 0 \quad \text{and} \quad \sum_{e^k \in D^k} \pi^k(e^k \mid t^k) = 1, \quad \forall d^k \in D^k, \quad \forall t^k \in T^k. \tag{6}$$

Since the payoffs for the agents of one principal may depend on the decisions of other principals and their agents, we cannot define incentive-compatibility for any one principal in isolation. Thus, we say that a mechanism π^k for corporation k is *incentive-compatible* in the context of mechanisms $(\pi^1, \ldots, \pi^N)$ iff it is an equilibrium for all of k's agents to be honest and obedient when principal k uses mechanism π^k, each other principal j uses the mechanism π^j, and the agents for every other principal j $(j \neq k)$ are expected to be honest and obedient. Formally, π^k is incentive-compatible for corporation k in the context of $(\pi^1, \ldots, \pi^N)$ iff

$$\sum_{\substack{t \in T \\ t_i^k = \tau_i^k}} \sum_{d \in D} P(t)\, \pi(d \mid t)\, U_i^k(d, t)$$

$$\geqq \sum_{\substack{t \in T \\ t_i^k = \tau_i^k}} \sum_{d \in D} P(t)\, \pi^{-k}(d^{-k} \mid t^{-k})\, \pi^k(d^k \mid t_{-i}^k, \hat{\tau}_i^k)\, U_i^k((d_{-k,i}, \delta_i^k(d_i^k)), t),$$

$$\forall i \in \{1, \ldots, n_k\}, \quad \forall \tau_i^k \in T_i^k, \quad \forall \hat{\tau}_i^k \in T_i^k, \quad \forall \delta_i^k: D_i^k \to D_i^k, \tag{7}$$

where

$$\pi(d \mid t) = \prod_{j=1}^{N} \pi^j(d^j \mid t^j) \quad \text{and} \quad \pi^{-k}(d^{-k} \mid t^{-k}) = \prod_{j \neq k} \pi^j(d^j \mid t^j).$$

[We use here the notation: $d=(d^1,\ldots,d^N)$, $t=(t^1,\ldots,t^N)$, $(t^k_{-i},\hat{\tau}^k_i)$ is the same as the vector t^k, except that t^k_i is replaced by $\hat{\tau}^k_i$, and $(d_{-k,i},\delta^k_i(d^k_i))$ differs from d in that the component d^i_k is changed to $\delta^k_i(d^k_i)$.] Inequality (7) asserts that k's ith agent should not be able to increase his expected utility above what he gets with honesty and obedience by reporting $\hat{\tau}^k_i$, when τ^k_i is his type, and by then choosing decision $\delta^k_i(d^k_i)$ when told to choose d^k_i.

When all agents are honest and obedient, principal k's expected utility from the mechanisms $(\pi^1,\ldots,\pi^N)$ is

$$V^k_0(\pi^1,\ldots,\pi^N)=\sum_{t\in T}\sum_{d\in D}P(t)\left(\prod_{j=1}^{N}\pi^j(d^j\mid t^j)\right)U^k_0(d,t). \tag{8}$$

Propositions 1 and 2 can be extended in the obvious way to the case of one principal designing a coordination mechanism in the context of fixed behavior by the other principals and their agents. That is, if $(\pi^1,\ldots,\pi^{k-1},\pi^{k+1},\ldots,\pi^N)$ characterizes the planned behavior in corporations other than k, then the optimal mechanism for principal k should maximize (8) over all $\pi^k: D^k\times T^k\to\boldsymbol{R}$ subject to the constraints (6) and (7). As before, this is a linear programming problem, but now principal k's optimal incentive-compatible mechanism depends on the mechanisms chosen by the other principals. Assuming that the principals act non-cooperatively in setting up their respective coordination mechanisms, we should then define $(\pi^1,\ldots,\pi^N)$ to be a *principals' equilibrium* iff, for each principal k, π^k maximizes (8) subject to (6) and (7), given the other π^j for all $j\neq k$. Unfortunately, we have the following result:

Proposition 3. Principals' equilibria do not always exist.

Proof. It suffices to show one example. So, consider a simple example where there are two principals ($N=2$), each with one agent ($n_k=1$). Each agent has type set $T^k_1=\{\alpha,\beta\}$, and the two agent's types are independent, $P(\alpha,\alpha)=P(\alpha,\beta)=P(\beta,\alpha)=P(\beta,\beta)=1/4$. Each principal has decision domain $D^k_0=\{A,B,C\}$, and each agent has no private decision ($|D^k_1|=1$). For $k=1$ or 2, the payoffs to principal k (U^k_0) and his agent (U^k_1) depend on the principal's decision and the agent's type according to the following matrix:

U^k_0, U^k_1	$t^k_1=\alpha$	$t^k_1=\beta$
$d^k_0=A$	$6,1$	$0,z^k$
$d^k_0=B$	$0,z^k$	$6,1$
$d^k_0=C$	$5,0$	$5,0$

Here, the term z^k is determined by the other principals actions as follows:

$$\text{for } k=1, \quad \begin{aligned} z^1 &= 2 \quad \text{if principal 2 chooses } A \text{ or } B, \\ &= 1 \quad \text{if principal 2 chooses } C, \end{aligned}$$

$$\text{for } k=2, \quad \begin{aligned} z^2 &= 2 \quad \text{if principal 1 chooses } C, \\ &= 1 \quad \text{if principal 1 chooses } A \text{ or } B. \end{aligned}$$

If $z^k=1$ for sure, then the optimal incentive-compatible mechanism for principal k is to choose A if the agent says his type is α, B if β. On the other hand, if there is any positive probability that $z^k=2$, then the agent will want to steer principal k to the 'wrong' corner (B if α, A if β), and the best incentive-compatible mechanism is to choose C for sure. So principal 1 wants to choose C if 2 might choose A or B with positive probability, and principal 2 wants to choose C if 1 might choose C, but principal 1 wants to choose A or B if 2 chooses C for sure, and principal 2 wants to choose A or B if 1 chooses A or B for sure. Thus, there cannot be any principals' equilibrium. Q.E.D.

Essentially, this non-existence result occurs because the set of incentive-compatible mechanisms for principal k varies upper-semicontinuously in the other π^j, rather than continuously as is required by the existence theorem of Debreu (1952). To obtain an existence theorem for the multi-principal problem, we must explore weaker solution concepts than full equilibrium. We shall develop one notion of *quasi-equilibrium* here. Our quasi-equilibria will differ from equilibria in that each principal's mechanism need not be optimal for him in this class of incentive-compatible mechanisms; instead, we only require that any mechanisms which he might prefer must be 'unsafe', in that they would not be incentive-compatible if the other principals made infinitesimal changes in their plans.

In our example, suppose that both principals planned to choose C. This is not a principal's equilibrium because there are better incentive-compatible mechanisms available to principal 1, if principal 2 is absolutely certain to choose C. However, if principal 1 (or his agent) believed that there was even an infinitesimal probability of principal 2 using A or B, then there would be no better incentive-compatible mechanism for principal 1 than to choose C for sure. Thus, we may refer to this pair of mechanisms (both principals choosing C for sure) as a quasi-equilibrium for the principals, since an infinitesimal perturbation in the mechanisms could render infeasible any mechanism which either principal might prefer to use.

Let $F^k(\pi^1,\ldots,\pi^N)$ be the set of all direct mechanisms $\tilde{\pi}^k$ for corporation k such that $\tilde{\pi}^k$ is incentive-compatible in the context of $(\pi^1,\ldots,\tilde{\pi}^k,\ldots,\pi^N)$. Let

$$W^k(\pi^1,\ldots,\pi^N)=\max_{\tilde{\pi}^k\in F^k(\pi^1,\ldots,\pi^N)} V_0^k(\pi^1,\ldots,\tilde{\pi}^k,\ldots,\pi^N).$$

That is, $W^k(\pi^1,\ldots,\pi^N)$ is the maximum expected expected utility which principal k can get from an incentive-compatible mechanism when the other principals use their π^j mechanisms. (Notice that F^k and W^k do not actually depend on the argument π^k.)

In general, we say that $(\bar{\pi}^1,\ldots,\bar{\pi}^N)$ is a *principals' quasi-equilibrium* iff there exists a sequence of direct mechanisms $\{(\pi_l^1,\ldots,\pi_l^N)\}_{l=1}^{\infty}$ such that, for every corporation k,

$$\bar{\pi}^k=\lim_{l\to\infty}\pi_l^k, \tag{9}$$

$$V_0^k(\bar{\pi}^1,\ldots,\bar{\pi}^N)\geqq\lim_{l\to\infty} W^k(\pi_l^1,\ldots,\pi_l^N), \tag{10}$$

$$\bar{\pi}^k\in F^k(\bar{\pi}^1,\ldots,\bar{\pi}^N). \tag{11}$$

Conditions (9) and (10) imply that, for any sequence $\{\hat{\pi}_l^k\}_{l=1}^{\infty}$ of mechanisms for corporation k, if $\hat{\pi}_l^k\in F^k(\pi_l^1,\ldots,\pi_l^N)$, $\forall l$, then

$$\begin{aligned}\limsup_{l\to\infty} V_0^k(\pi_l^1,\ldots,\hat{\pi}_l^k,\ldots,\pi_l^N)&\leqq\lim_{l\to\infty} W^k(\pi_l^1,\ldots,\pi_l^N)\\&\leqq V_0^k(\bar{\pi}^1,\ldots,\bar{\pi}^N)\\&=\lim_{l\to\infty} V_0^k(\pi_l^1,\ldots,\pi_l^k,\ldots,\pi_l^N).\end{aligned}$$

So (9) and (10) imply that the sequence of mechanisms $\{(\pi_l^1,\ldots,\pi_l^N)\}_{l=1}^{\infty}$ is asymptotically optimal for every principal, and (11) implies that the sequence is also asymptotically incentive-compatible for every corporation. The limit of any such sequence is a quasi-equilibrium.

To put it differently, (9) and (10) imply that, for any direct mechanism $\hat{\pi}^k$, if $V_0^k(\bar{\pi}^1,\ldots,\hat{\pi}^k,\ldots,\bar{\pi}^N)>V_0^k(\bar{\pi}^1,\ldots,\bar{\pi}^k,\ldots,\bar{\pi}^N)$, then

$$\hat{\pi}^k\notin F^k(\pi_l^1,\ldots,\pi_l^N)$$

for all sufficiently large l. That is, we can make arbitrarily small perturbations in our quasi-equilibrium (from $\bar{\pi}^k$ to π_l^k) in such a way as to render infeasible any mechanism which a principal might have preferred to use.

It should be clear that (letting $\pi_l^k=\bar{\pi}^k$) any principals' equilibrium is also a

principals' quasi-equilibrium. However, the quasi-equilibria are a larger set, and do always exist.

Proposition 4. A quasi-equilibrium must exist, for any multi-principal problem with finite type sets and finite decision sets.

Proof. Given any $\varepsilon>0$, we say that π^k is *ε-incentive-compatible* for corporation k in the context of $(\pi^1,\ldots,\pi^N)$ iff π^k is a direct mechanism and violates none of the incentive constraints (7) for corporation k by more than ε; that is,

$$\sum_{\substack{t\in T\\ t_i^k=\tau_i^k}}\sum_{d\in D}[P(t)\,\pi(d\,|\,t)\,U_i^k(d,t)]+\varepsilon$$

$$\geqq \sum_{\substack{t\in T\\ t_i^k=\tau_i^k}}\sum_{d\in D} P(t)\,\pi^{-k}(d^{-k}\,|\,t^{-k})\,\pi^k(d^k\,|\,t^k_{-i},\hat{\tau}_i^k)\,U_i^k((d_{-k,i},\delta_i^k(d_i^k)),t),$$

$$\forall i\in\{1,\ldots,n_k\},\quad \forall\tau_i^k\in T_i^k,\quad \forall\hat{\tau}_i^k\in T_i^k,\quad \forall\delta_i^k\colon D_i^k\to D_i^k. \tag{12}$$

We let $\bar{F}^k(\pi^1,\ldots,\pi^N,\varepsilon)$ be the set of all direct mechanisms $\tilde{\pi}^k$ such that $\tilde{\pi}^k$ is ε-incentive-compatible for corporation k in the context of $(\pi^1,\ldots,\tilde{\pi}^k,\ldots,\pi^N)$.

It is straightforward to show that $\bar{F}^k(\cdot,\varepsilon)$ is an upper-semicontinuous correspondence, and that $\bar{F}^k(\pi^1,\ldots,\pi^N,\varepsilon)$ is always a non-empty compact convex set. The key step is to show that, if $\varepsilon>0$, then $\bar{F}^k(\cdot,\varepsilon)$ is also lower-semicontinuous. [See Debreu (1959, ch. 1).] To check lower-semicontinuity, let $\hat{\pi}^k$ be any direct mechanism for corporation k such that $\hat{\pi}^k(d^k\,|\,t^k)$ is independent of t^k. Then, for any π^k in $\bar{F}^k(\pi^1,\ldots,\pi^N,\varepsilon)$ and any $0<\lambda<1$, $(1-\lambda)\pi^k+\lambda\hat{\pi}^k$ will satisfy the constraints in (12) strictly, and so $(1-\lambda)\pi^k+\hat{\pi}^k$ will be in $\bar{F}^k(\tilde{\pi}^1,\ldots,\tilde{\pi}^N,\varepsilon)$ for all $(\tilde{\pi}^1,\ldots,\tilde{\pi}^N)$ sufficiently close to $(\pi^1,\ldots,\pi^N)$. This is sufficient to prove that $\bar{F}^k(\cdot,\varepsilon)$ is lower-semicontinuous, and thus continuous, for any positive ε.

Let $G^k(\pi^1,\ldots,\pi^N,\varepsilon)$ be the set of mechanisms $\tilde{\pi}^k$ which maximize $V_0^k(\pi^1,\ldots,\tilde{\pi}^k,\ldots,\pi^N)$ subject to $\tilde{\pi}^k\in\bar{F}^k(\pi^1,\ldots,\pi^N,\varepsilon)$. It is straightforward to check that $G^k(\pi^1,\ldots,\pi^N,\varepsilon)$ is a non-empty compact convex subset of $\bar{F}^k(\pi^1,\ldots,\pi^N,\varepsilon)$; and $G^k(\cdot,\varepsilon)$ is upper-semicontinuous, because $\bar{F}^k(\cdot,\varepsilon)$ is continuous [by the Maximum Theorem of Berge (1959), see Debreu (1959, p. 19)]. By the Kakutani Fixed-Point Theorem, for any $l>0$, there exists some $(\pi_l^1,\ldots,\pi_l^N)$, such that $\pi_l^k\in G^k(\pi_l^1,1/l)$ for every k. These mechanisms satisfy $\pi_l^k\in\bar{F}^k(\pi_l^1,\ldots,\pi_l^N,1/l)$, and $V_0^k(\pi_l^1,\ldots,\pi_l^N)\geqq W^k(\pi_l^1,\ldots,\pi_l^N)$ because $\bar{F}^k(\pi_l^1,\ldots,\pi_l^N,1/l)\supseteq F^k(\pi_l^1,\ldots,\pi_l^N)$.

The set of all direct mechanisms is compact, so we can select a convergent subsequence of the $\{(\pi_l^1,\ldots,\pi_l^N)\}$, converging to some $(\bar{\pi}^1,\ldots,\bar{\pi}^N)$ as $l\to\infty$. As $l\to\infty$ and $\varepsilon=1/l\to 0$, the ε-incentive-compatibility constraints (12) converge to

the incentive-compatibility contraints (7). So $\bar{\pi}^k \in F^k(\bar{\pi}^1, \ldots, \bar{\pi}^N)$ and

$$V_0^k(\bar{\pi}^1, \ldots, \bar{\pi}^N) = \lim_{l \to \infty} V_0^k(\pi_l^1, \ldots, \pi_l^N)$$

$$\geqq \lim_{l \to \infty} W^k(\pi_l^1, \ldots, \pi_l^N).$$

Thus $(\bar{\pi}^1, \ldots, \bar{\pi}^N)$ is a principals' quasi-equilibrium. Q.E.D.

References

Antle, R., 1980, Moral hazard in auditor contracts, Ph.D. dissertation (Stanford University, Stanford, CA).

Aumann, R.J., 1974, Subjectivity and correlation in randomized strategies, Journal of Mathematical Economics 1, 67–96.

Dasgupta, P.S., P.J. Hammond and E.S. Maskin, 1979, The implementation of social choice rules: Some results on incentive compatibility, Review of Economic Studies 46, 185–216.

D'Aspremont, C. and L.-A. Gérard-Varet, 1979, Incentives and incomplete information, Journal of Public Economics 11, 25–45.

Debreu, G., 1952, A social existence theorem, Proceedings of the National Academy of Sciences of the USA 38, 886–893.

Debreu, G. 1959, Theory of value (Yale University Press, New Haven, CT).

Gibbard, A., 1973, Manipulation of voting schemes: A general result, Econometrica 41, 587–602.

Harris, M. and A. Raviv, 1979, Optimal incentive contracts with imperfect information, Journal of Economic Theory 20, 231–259.

Harris, M. and R.M. Townsend, 1981, Resource allocation under asymmetric information, Econometrica 49, 33–64.

Harsanyi, J.C., 1967/8, Games with incomplete information played by 'Bayesian' players, Management Science 14, 159–189, 320–334, 486–502.

Holmström, B., 1977, On incentives and control in organizations, Ph.D. dissertation (Stanford University, Stanford, CA).

Holmström, B., 1979, Moral hazard and observability, Bell Journal of Economics 10, 74–91.

Holmström, B., 1980, On the theory of delegation, Mimeo. (Northwestern University, Evanston, IL); to appear in: M. Boyer and R. Kihlstrom, Bayesian models in economic theory.

Mirrlees, J., 1976, The optimal structure of incentives and authority within an organization, Bell Journal of Economics 7, 105–131.

Myerson, R.B., 1979, Incentive compatibility and the bargaining problem, Econometrica 47, 61–73.

Myerson, R.B., 1981, Optimal auction design, Mathematics of Operations Research 6, 58–73.

Rosenthal, R.W., 1978, Arbitration of two-party disputes under uncertainty, Review of Economic Studies 45, 595–604.

Ross, S., 1973, The economic theory of agency: The principal's problem, American Economic Review 63, 134–139.