INCENTIVE CONSTRAINTS AND OPTIMAL COMMUNICATION SYSTEMS

Roger B. Myerson
J. L. Kellogg Graduate School of Management
Northwestern University
Evanston, IL 60208

ABSTRACT

An example of a sender-receiver game, due to Farrell, is studied to illustrate the role of incentive constraints in the design of optimal communication systems between rational individuals whose interests are not the same. Some noise in the communication system is essential for substantive communication. Informational and strategic incentive constraints are linear, so finding an optimal incentive-compatible mediation plan is a linear programming problem. The revelation principle guarantees that an optimal incentive-compatible mediation plan is also optimal among all equilibria with all possible communication systems. Without moral hazard, participational incentive constraints replace strategic incentive constraints.

1. An introductory example.

The ability of people to communicate is often limited by differences in their interests and incentives, which may prevent them from trusting one another. People cannot be expected to testify against themselves, nor can they be expected to exert efforts for which they will not be rewarded. Thus, when one person says something about his private information or makes some promise about a decision that he will make, these statements may not be believable if they contradict the person's incentives. That is, there are incentive constraint that limit communication in a very different way from the channel capacities that have been traditionally studied by information theorists and engineers. Game theory provides the foundation for the mathematical theory of optimal design of communication systems subject to incentive constraints. This paper offers an introduction to this theory, in the context of a simple example due to Farrell [1986]. For a broader introduction to this subject, see also Aumann [1974, 1987], Forges [1986], and Myerson [1985, 1986].

We can set Farrell's example in a realistic situation using the following story. An saleswoman who works for a large company has recently given birth to her first child. Her long-term goal may be either to *remain* with the company for many years, or to *quit* within a year or so. She knows her long-term goal, but her boss does not. Under a recent company directive, her boss must make some cuts in his sales staff. After these cuts, some salespeople will have to take on additional responsibilities that will involve a promotion. If the long-term goal of the saleswoman in question were to remain with the company, then her boss would prefer to give her additional responsibilities and a promotion. On the other hand, if her long-term goal were to quit within a year, then the boss would prefer to fire her now, as a part of his required cuts.

Let us denote the boss's three options, with regard to this saleswoman, as follows:

$\underline{A}$ = "give her additional responsibilities and a promotion,"
$\underline{B}$ = "neither fire nor promote her"
$\underline{C}$ = "fire her to achieve the required cuts in staff,"

To describe the saleswoman's information, let us say that her type is $\underline{R}$ if her long-term goal is to remain with the firm, and that her type is $\underline{Q}$ if her goal is to quit within about a year. Let us suppose that her boss currently thinks that these two possible types are equally likely to be true; that is, his subjective probabilities are

$$P(R) = P(Q) = 0.5 .$$

To describe the incentives of the saleswoman and her boss, let us suppose that their payoffs (measured in some utility scale, as defined by von Neumann and

Morgenstern) depend on her type and his action as follows:

	Boss's Action:		
Saleswoman's Type:	A	B	C
R	8	4	0
Q	1	2	0

SALESWOMAN'S PAYOFF

	Boss's Action:		
Saleswoman's Type:	A	B	C
R	3	2	0
Q	-3	-1	0

BOSS'S PAYOFF

All of the above information is assumed to be common knowledge among the saleswoman and her boss. In addition, the saleswoman has one additional piece of private information: she knows her true type (R or Q).

Crawford and Sobel [1982] began the study of sender-receiver games. In any sender-receiver game, there are two players, the "sender" and the "receiver." The sender has private information but has no substantive payoff-relevant actions. The receiver has a range of payoff-relevant actions to choose among but has no private information. Our example is a sender-receiver game, where the saleswoman is the sender and the boss is the receiver. As the terminology suggests, we assume that the sender can send messages about her private information to the receiver, through some communication system, to try to influence his action. Our basic question is, what kinds of communication systems are best for the players in this game?

2. Failure of direct communication.

Farrell [1986] has shown that in this example, if the saleswoman and her boss communicate directly and if they both behave rationally and intelligently, then the boss cannot be influenced by the saleswoman's words; he must choose option B (neither fire nor promote) for any message that she might communicate with positive probability. We now review the proof of this result.

The boss initially assigns probability .5 to each of the saleswoman two possible types. However, after getting a message from the saleswoman, the boss might revise his beliefs to assign some other probability ρ to the event that the saleswoman's type is R (planning to remain), and to assign probability $1 - \rho$ to the event that the saleswoman's type is Q (planning to quit). The optimal action for the boss would be the option that maximizes the boss's expected payoff, which depends on this number ρ. If the boss chose action A (promote her), his expected payoff would be

$$3\,\rho + -3\,(1 - \rho) = -3 + 6\,\rho.$$

If the boss chose action B (neither promote nor fire her), his expected payoff would be

$$2\,\rho + -1\,(1 - \rho) = -1 + 3\,\rho.$$

If the boss chose action C (fire her), his expected payoff would be

$$0\,\rho + 0\,(1 - \rho) = 0.$$

Thus, action A is optimal for the boss when both $-3 + 6\,\rho \geq -1 + 3\,\rho$ and $-3 + 6\,\rho \geq 0$, which happens if and only if

$$\rho \geq 2/3.$$

Similarly, action B is optimal when both $-1 + 3\,\rho \geq -3 + 6\,\rho$ and $-1 + 3\,\rho \geq 0$, which happens if and only if

$$2/3 \geq \rho \geq 1/3.$$

Finally, action C is optimal when both $0 \geq -3 + 6\,\rho$ and $0 \geq -1 + 3\,\rho$, which happens if and only if

$$1/3 \geq \rho.$$

If $\rho = 2/3$, the boss would be indifferent between actions A and B, and he would be willing to randomize between these two actions. Similarly, if $\rho = 1/3$, the boss would be indifferent between actions B and C, and he would be willing to randomize between these two actions.

Notice that there are no beliefs that leave the boss willing to randomize between actions A and C, or between all three actions. That is, no matter what message the saleswoman might convey to her boss, there must be some number β such that $0 \leq \beta \leq 1$ and his expected rational response is either

(1) to randomize between actions A and B, choosing B with probability β and choosing A with probability $1 - \beta$; or

(2) to randomize between actions B and C, choosing B with probability β and choosing C with probability $1 - \beta$.

Now, suppose that the saleswoman's type is actually R. Then the saleswoman cannot be indifferent between any of two of these possible responses, because she prefers A over B and prefers B over C. That is, if one message would elicit a response in case (1) and another would elicit a response in case (2), she would strictly prefer to send the message that elicits the response in case (1). If two messages would elicit different responses in case (1), she would strictly prefer the message that elicits the lower probability of B. If two messages would elicit different responses in case (2), she would strictly prefer the message that elicits the higher probability of B. Thus, among all messages that the saleswoman could send to her boss, the messages that she would most prefer to send when her type is R must all elicit exactly the same response from her boss. Since they all elicit the same response, there is no point in distinguishing between these messages (the boss himself does not, in effect). So without loss of generality we can assume that there is a unique message that the saleswoman would most want to send to her boss if her type were R, when the boss's expected response to every possible message are taken into account. We may call this message, "my type is R."

When two people communicate directly, with no possibility of noise or distortion in their communication system, the message that one individual receives from another rational individual must be the message that the latter most prefers to send. Thus, when communication is direct there can essentially be only one message sent (with probability 1) by the saleswoman to her boss if her type is R.

Now, let δ denote the conditional probability that the saleswoman would send this same message ("my type is R") if her type were Q. Then, by Bayes' rule, the conditional probability that the boss would assign to her type being R after receiving this message would be

$$\rho = (.5 \times 1)/(.5 \times 1 + .5 \times \delta) = 1/(1 + \delta) \geq .5$$

(because $\delta \leq 1$), so the boss's rational response to this message must be either B, or A, or some randomization between A and B. If there were any other message that the saleswoman would send with positive probability when her type was Q, then the boss would be sure that her type was Q after receiving this other message, so his rational response to this other message would be to choose C (because C is his best action when he is sure that her type is Q). But, if her type was Q, the saleswoman would prefer B or A or any randomization between A and B over getting C for sure. Thus, there could not be any other message that she would want to send with positive probability when her type was Q. That is, if her type was Q, then, in direct communication, the saleswoman would rationally send the same message ("my type is R") as she would send if her type was R.

What has just been shown is that both types must always send essentially the same message, so that the boss can always ignore what the saleswoman says and do what was optimal for him before he heard her message. Since he had $P(R) = .5$ before getting any message, and since B is the best action for him when $\rho = P(R) = .5$, his optimal response must be to choose B (neither fire nor promote) after anything that the saleswoman might say to him with positive probability. That is, informative communication from the saleswoman to her boss is impossible when there is no noise in the communication system.

This conclusion is remarkable because informative communication is possible through noisy channels. For example, suppose that there is some third party, say a former supervisor of this saleswoman, who relays messages only with probability 0.4 . Consider the following scenario. If the saleswoman's type were R, she would tell her former supervisor that she plans to remain and hopes to get the promotion (boss's action A), in which case he would relay this message to her current boss with probability .4, and would relay no message (silence) with probability .6 . On the other hand, if the saleswoman's type were Q, she would say nothing to her former supervisor and he would be sure to relay no message to her current boss. If the boss got the relayed message from the former supervisor, then the boss would be sure that the saleswoman's type must be R, and so he would choose action A (his rational move when $\rho = 1$). If the boss got no relayed message, then by Bayes' rule he would update his subjective probability of the saleswoman's type being R to

$$\rho = (.5 \times .6)/(.5 \times .6 + .5 \times 1) = .375,$$

so he would choose action B (his rational move when $2/3 \geq \rho \geq 1/3$).

Under this scenario, if the saleswoman's type were R then she would get an expected payoff of $.4 \times 8 + .6 \times 4 = 5.6$ from speaking to her former supervisor, which would be better than the expected payoff of 4 (from action B) that she would get from not speaking to him. On the other hand, if the saleswoman's type were Q then she would get an expected payoff of $.4 \times 1 + .6 \times 2 = 1.6$ from speaking to the former supervisor (since there would be a 40% chance that this would lead her to getting a promotion that she does not want) which would be worse than the payoff of 2 (from action B) that she would get under this scenario from not speaking to the former supervisor. Thus, it is indeed rational for the saleswoman to speak to her former supervisor if and only if her type is R, in this scenario. That is, the scenario is a rational equilibrium for the saleswoman and her boss.

The moral of this story is that a noisy imperfect communication system (such as a third party who sometimes fails to pass on a message) may in some situations actually improve the possibilities for meaningful communication between rational individuals. In this case, the imperfection of the former supervisor as a communication channel helps the saleswoman to communicate meaningfully, because the possibility of his not relaying the message guarantees that the current boss could not infer, from the absence of any relayed message, that the saleswoman did not speak to her former supervisor about her desire for the promotion. Thus, the saleswoman of type Q could refrain from speaking to her former supervisor without fearing that her boss will infer that she is planning to quit soon and therefore fire her now. In effect, the noise in the communication system gives a kind of protection to the saleswoman if her type is Q and she does not send the same message as she would have send if her type were R.

The mediated communication system described above section is not the best possible communication system for the saleswoman, even though it is better than face-to-face communication. It is straightforward to check that if the former supervisor's probability of relaying the saleswoman's message was increased up to .5, it would still be a rational equilibrium for her to speak to the former supervisor if and only if her type is R, and for the boss to choose action A if he gets the relayed message from the former supervisor and to choose action B if he gets no relayed message.

However, if the probability of relaying the message were higher than .5, then this equilibrium would break down. To see why, let θ denote the probability that the former supervisor would relay the saleswoman's message. By Bayes' rule, if the boss did <u>not</u> get a relayed message, then the conditional probability that he would assign to the event that the saleswoman's goal is to remain would be

$$\rho = .5\times(1 - \theta)/(.5 \times (1 - \theta) + .5 \times 1) = (1 - \theta)/(2 - \theta).$$

If $\theta > .5$, then $\rho < 1/3$, and so the boss's rational response would be action C (to fire the saleswoman) if he got no relayed message. But if the saleswoman expected this response, she would ask her former supervisor to relay the message even if her type were Q, so the relayed message would no longer convey any information.

3. Optimal incentive-compatible mediation plans.

The preceding analysis begs the question, can we design an even better communication system or mediation plan for helping the saleswoman to communicate with her boss in this example? How can we identify optimal communication systems? Once we admit that it is not necessarily optimal to simply maximize channel capacity and minimize noise in a communication system, it may at first seem difficult to know how a mediator can best help two rational individuals to communicate effectively.

To identify optimal communication systems, we first consider what may seem to be a more restricted class of communication systems, called incentive-compatible mediation plans.

Consider a general situation involving two or more individuals, some of whom have private information unknown to others, and some of whom have a range of actions to choose among. Suppose that a mediator wants to help these individuals to communicate with each other, so that individuals' actions can be correlated with each other and can depend on each other's information in a way that may make some or all of them all better off. Let us suppose that the mediator will operate as follows. First he will ask each individual to independently and confidentially report the state of his or her private information, that is, his or her type, to the mediator. Then, after collecting these reports, the mediator will confidentially recommend to each individual which action he or she should choose, among the actions available to him. A mediation plan is any rule that specifies how the mediator will determine the actions that he recommends, as a (possibly random) function of the reports that he receives. Mathematically, a mediation plan is defined by specifying, for every possible combination of type-reports that the mediator might receive, and for every possible combination of actions that he might recommend, what is the conditional probability of his recommending these actions if he gets these reports.

In our example, the mediator could get one of two possible reports, "R" or "Q," from the saleswoman, and he could send one of three possible recommendations to the boss, "A" or "B" or "C." Let us denote the conditional probabilities that make up a mediation plan by "$\mu(\cdot|\cdot)$." That is, we let $\mu(A|R)$ denote the conditional probability that the mediator would recommend action A to the boss if the saleswoman reported that her type was R, and so on. To fully specify a mediation plan, we need to specify six numbers: $\mu(A|R)$, $\mu(B|R)$, $\mu(C|R)$, $\mu(A|Q)$, $\mu(B|Q)$, and $\mu(C|Q)$. Since these numbers are supposed to represent probabilities, they must all be nonnegative. Furthermore, they must satisfy the following two probability constraints:

$$\mu(A|R) + \mu(B|R) + \mu(C|R) = 1,$$
$$\mu(A|Q) + \mu(B|Q) + \mu(C|Q) = 1.$$

These constraints assert that, given either report "R" or "Q" from the saleswoman, the conditional probabilities of the mediator recommending each of the three possible actions must sum to one.

An incentive-compatible mediation plan is a plan such that, if each individual expects that the other individuals will report their types honestly to the mediator and obey his recommendations, then no individual could expect to gain by lying to the mediator or by disobeying his recommendations. That is, an incentive-compatible mediation plan is one such that it is an equilibrium for all individuals to be honest and obedient to the mediator.

The set of incentive-compatible mediation plans can be characterized by some mathematical inequalities called incentive constraints, which express mathematically the requirement that the mediation plan should not give any individual an incentive to be dishonest or disobedient, under any possible circumstances. We may distinguish two kinds of incentive constraints: informational incentive constraints, which express the requirement that no individual should have any incentive to lie to the mediator; and strategic incentive constraints, which express the requirement that no individual should have any incentive to disobey the mediator.

In this example, there are two informational incentive constraints. One informational incentive constraint asserts that the saleswoman should not expect to gain by reporting that her type is R (planning to remain) if her type is actually Q (planning to quit). This constraint may be written as follows:

$$1\ \mu(A|Q) + 2\ \mu(B|Q) + 0\ \mu(C|Q) \geq 1\ \mu(A|R) + 2\ \mu(B|R) + 0\ \mu(C|R).$$

The left-hand side of this constraint is the expected payoff to the saleswoman under the mediation plan if her type is Q and she is honest, and the right-hand side of this constraint is the expected payoff to the saleswoman under the mediation plan if her type is Q but she lies and reports "R" as her type. The constraint asserts that her expected payoff from honesty is not less than her expected payoff from lying, when her actual type is Q.

The other informational incentive constraint asserts that the saleswoman should not expect to gain by reporting that her type is Q if her type is actually R. This constraint may be written as follows:

$$8\ \mu(A|R) + 4\ \mu(B|R) + 0\ \mu(C|R) \geq 8\ \mu(A|Q) + 4\ \mu(B|Q) + 0\ \mu(C|Q).$$

If the mediation plan satisfies both of these informational incentive constraints, then the saleswoman will never have any incentive to lie to the mediator, as long as the boss is expected to obey the recommendations.

There are six strategic incentive constraints for our example. These constraints assert that, for each action of the three actions that the mediator might recommend to the boss, he would not expect to gain by disobediently choosing one of the other two actions instead. For example, let us formulate the constraint which asserts that the boss would not expect to gain by choosing action B (neither fire nor promote her) if the mediator recommends action A (promote her). First we must derive the boss's beliefs about the saleswoman if he gets this recommendation. Recall that we are assuming that the saleswoman's report to the mediator was made confidentially, so that the boss does not know what she actually reported. But we assume he does understand the mediation plan, so he can make some Bayesian inference about her type from the fact that the mediator has recommended action A to him. Since he assigned a prior probability of .5 to each of the possible types before learning the mediator's recommendation, the posterior probability that the boss should assign to the event that the saleswoman's type is R, after learning that the mediator recommends action A, is

$$\rho = .5 \times \mu(A|R)/(.5 \times \mu(A|R) + .5 \times \mu(A|Q)) = \mu(A|R)/(\mu(A|R) + \mu(A|Q)).$$

Similarly, the posterior probability that he should assign to type Q after getting recommendation "A" is

$$(1 - \rho) = \mu(A|Q)/(\mu(A|R) + \mu(A|Q)).$$

With these beliefs, if the boss obeyed the recommendation to choose A, his expected payoff would be

$$3\ \rho + -3\ (1 - \rho) = (3\ \mu(A|R) + -3\ \mu(A|Q))/(\mu(A|R) + \mu(A|Q));$$

but if he disobediently chose action B, his expected payoff would be

$$2\ \rho + -1\ (1 - \rho) = (2\ \mu(A|R) + -1\ \mu(A|Q))/(\mu(A|R) + \mu(A|Q)).$$

Thus, to not give any incentive to choose B when A is recommended, the mediation plan must satisfy

$$\begin{aligned}&(3\ \mu(A|R) + -3\ \mu(A|Q))/(\mu(A|R) + \mu(A|Q))\\ &\qquad \geq (2\ \mu(A|R) + -1\ \mu(A|Q))/(\mu(A|R) + \mu(A|Q)).\end{aligned}$$

We can multiply both sides of this constraint by the common denominator $(\mu(A|R) + \mu(A|Q))$, to get

$$3\ \mu(A|R) + -3\ \mu(A|Q) \geq 2\ \mu(A|R) + -1\ \mu(A|Q),$$

or, more simply,

$$1\ \mu(A|R) - 2\ \mu(A|Q) \geq 0.$$

This inequality is the strategic incentive constraint which asserts that the boss should have no incentive for choosing B when A is recommended. The other five strategic incentive constraints may similarly derived and are as follows:

$3\ \mu(A|R) - 3\ \mu(A|Q) \geq 0$	(no incentive for C when A is recommended)
$-1\ \mu(B|R) + 2\ \mu(B|Q) \geq 0$	(no incentive for A when B is recommended)
$2\ \mu(B|R) - 1\ \mu(B|Q) \geq 0$	(no incentive for C when B is recommended)
$-3\ \mu(C|R) + 3\ \mu(C|Q) \geq 0$	(no incentive for A when C is recommended)
$-2\ \mu(C|R) + 1\ \mu(C|Q) \geq 0$	(no incentive for B when C is recommended)

If a mediation plan satisfies all six of these strategic incentive constraints, then the boss will never be tempted to disobey the mediator, if the saleswoman is expected to be honest to the mediator.

Thus an incentive-compatible mediation plan must satisfy two probability constraints, two informational incentive constraints, and six strategic incentive constraints. Notice that all of these constraints depend linearly on the various components of the mediation plan $\mu(\bullet|\bullet)$.

Now, suppose that we want to find the incentive-compatible mediation plan that maximizes the expected payoff to the saleswoman when her type is R. With type R, her expected payoff would be

$$8\ \mu(A|R) + 4\ \mu(B|R) + 0\ \mu(C|R).$$

Notice that this expected payoff is also linear in $\mu(\cdot|\cdot)$. Thus, our problem is to maximize a linear function of μ, subject to ten linear constraints. This problem is an example of a linear programming problem, which can be solved efficiently by many widely-available computer programs. The unique optimal solution to this problem is

$$\mu(A|R) = .8,\quad \mu(B|R) = .2,\quad \mu(C|R) = 0,$$
$$\mu(A|Q) = .4,\quad \mu(B|Q) = .4,\quad \mu(C|Q) = .2\ .$$

This mediation plan gives the saleswoman an expected payoff of 7.2 when her type is R, which is higher than she could expect with type R under any other incentive-compatible mediation plan.

4. The revelation principle.

It is now natural to ask, why should we restrict our attention to incentive-compatible mediation plans? If we consider other mediation plans or communication systems that give some incentive to be dishonest or disobedient, is it possible that the expected outcome might be better (according to whatever criterion we have in mind) than the best incentive-compatible mediation plan?

Notice that any communication system creates a game in which the individuals choose strategies for sending messages as a function of their private information, and for ultimately choosing their payoff-relevant actions as a function of their private information and the messages that they receive. A theorem known as the revelation principle asserts that, for any communication system and any equilibrium of this communication game, there is an equivalent incentive-compatible mediation plan that gives all types of all individuals the same expected payoff. Thus, by the revelation principle, the highest expected payoff that an individual can expect in an incentive-compatible mediation plan is also the highest that he can expect in any equilibrium of the game generated by any communication system. That is, the answer to the above question is No, if the individuals are assumed to behave rationally and intelligently. So an optimal incentive-compatible mediation plan is also optimal among all possible equilibria of all possible communication systems.

To prove the revelation principle, suppose that someone has given us a proposed communication system or mediation plan that is not incentive compatible. Suppose also that this person has given us a description of the strategies that the individuals would be expected to use, to determine their reports and actions with this communication system. We can construct an equivalent incentive-compatible mediation plan, by instructing a mediator to behave as follows. First, the mediator should ask every individual to report his or her type to the mediator confidentially and independently. Second, the mediator should compute the input messages that each individual would send into the given communication system, according to the given strategies, if their types were as reported. Third, the mediator should compute the output messages that each individual would receive from the given communication system, if the computed input messages were sent. Fourth, the mediator should compute the actions that each individual would choose, according to the given strategies, if the computed output messages were received and if their types were as initially reported. Finally, the mediator should recommend confidentially to each individual that he should choose the action just computed for him.

A mediator who behaves in this way is effectively simulating the given communication system and the given communication strategies, so the constructed mediation plan does give each type of each individual the same expected payoff as under the given communication system and strategies. Furthermore, if the given strategies form an equilibrium, then the constructed mediation plan must be incentive compatible. If it were not, then there would some individual who could gain by lying to the mediator or disobeying him. But then, since our mediator is effectively just simulating this individual's given strategy in the given communication system, this individual could have gained, in the context of the given communication system, by lying to himself before implementing his own strategy, or by disobeying the recommendation that his own strategy generates for him. Of course, such a conclusion is impossible if the given strategies form a rational equilibrium.

To illustrate this argument in the context of our example, consider a communication system in which the saleswoman can either send the message "my type is R" or be silent (send the message " "). If she sends the message "my type is R," then the boss will receive the message "her type is R" with probability θ, otherwise her boss will receive only silence (the message " "). If $\theta \leq .5$ then, as we have seen, the following pair of strategies form an equilibrium with this communication system: the saleswoman sends the message "my type is R" if and only if her type actually is R; and the boss chooses action A if he receives the message "her type is R," whereas he chooses action B if he hears only silence from the communication system. This communication system with this pair of equilibrium strategies is clearly equivalent to the following mediation plan:

$$\mu(A|R) = \theta, \quad \mu(B|R) = 1 - \theta, \quad \mu(C|R) = 0,$$
$$\mu(A|Q) = 0, \quad \mu(B|Q) = 1, \quad \mu(C|Q) = 0,$$

It is straightforward to check that this mediation plan does satisfy all of the incentive constraints as long as $\theta \leq .5$.

Both types of the saleswoman would get their most-preferred outcomes if the above strategies could be rationally applied with to this communication system when $\theta = 1$. However, if $\theta > .5$, then there is a change in the equilibrium strategies that describe how the saleswoman and her boss might rationally behave with this communication system. A pair of equilibrium strategies when $\theta > .5$ is as follows: the saleswoman sends the message "my type is R" for sure, no matter what her actual type is; and the boss always chooses action B, whether he hears "her type is R" or silence. The mediation plan that is equivalent to the given communication system with this equilibrium is just

$$\mu(A|R) = 0, \quad \mu(B|R) = 1, \quad \mu(C|R) = 0,$$
$$\mu(A|Q) = 0, \quad \mu(B|Q) = 1, \quad \mu(C|Q) = 0,$$

which is trivially incentive compatible.

5. Participational incentive constraints without moral hazard.

Thus far, we have assumed that the boss has inalienable control over his choice among the three possible actions, so that he cannot precommit himself to any strategy that he might regret or want to revise at the time when he actually implements the action that the strategy designates. This inability to precommit to a strategy is called moral hazard.

To appreciate the importance of moral hazard, let us consider the problem of designing the mediation plan that is best for the boss in our example. The boss's expected payoff from a mediation plan $\mu(\bullet|\bullet)$ is

$$.5 \times (3\ \mu(A|R) + 2\ \mu(B|R) + 0\ \mu(C|R)) + .5 \times (-3\ \mu(A|Q) - 1\ \mu(B|Q) + 0\ \mu(C|Q)).$$

The following mediation plan maximizes this expected payoff subject to the probability constraints and incentive constraints that we listed in section 3:

$$\mu(A|R) = 2/3, \quad \mu(B|R) = 1/3, \quad \mu(C|R) = 0,$$
$$\mu(A|Q) = 0, \quad \mu(B|Q) = 2/3, \quad \mu(C|Q) = 1/3.$$

The expected payoff to the boss in this plan is equal to 1.

Consider, however, the following plan:

$$\mu(A|R) = 1, \quad \mu(B|R) = 0, \quad \mu(C|R) = 0,$$
$$\mu(A|Q) = 0, \quad \mu(B|Q) = .5, \quad \mu(C|Q) = .5\ .$$

This plan offers the boss a higher expected payoff of 1.25, and it satisfies all of the probability constraints and informational incentive constraints from section 3. However, it violates one strategic incentive constraint: the constraint that the boss should have no incentive to choose action C when action B is recommended ($2\ \mu(B|R) - 1\ \mu(B|Q) \geq 0$). That is, this mediation plan is infeasible, because the assumption of moral hazard makes it impossible to believe that the boss would actually choose action B when the mediator recommended it under this plan. The problem is that, even though the boss would prefer this plan, the saleswoman would not trust him to obey it.

However, there are some situations in which individuals can make binding commitments to obey a mediator's recommendations. That is, individuals may have the option to voluntarily accept binding mediation, in which they would subsequently have no choice but to implement the mediator's recommendations. In such situations, we say that there is no moral hazard. In situations with no moral hazard, the strategic incentive constraints must be dropped from the definition of incentive compatibility and, in their place, we must add some <u>participational incentive constraints</u>. These participational incentive constraints (or individual rationality constraints, as they are often called), assert that each type of each individual should be willing to make a binding commitment to obey the recommendations of the mediation plan. To achieve this willingness, the mediation plan must offer each type of each individual an expected payoff that is not less than the best expected payoff that he could guarantee himself without any cooperation from anyone else.

To formulate participational incentive constraints, we need to determine what each individual could expect or guarantee himself if he did not agree to commit himself to obey the mediator. In our example, the highest expected payoff that the boss could guarantee himself without any mediation is equal to .5 . He can achieve this payoff by simply choosing action B without any message from the saleswoman, which gives him the expected payoff of $.5 \times 2 + .5 \times -1 = .5$. On the other hand, the saleswoman cannot guarantee herself an expected payoff higher than 0, since she cannot prevent her boss from firing her. Thus, the participational incentive constraints for this example would be

$$\begin{aligned} &.5 \times (3\ \mu(A|R) + 2\ \mu(B|R) + 0\ \mu(C|R)) \\ &\qquad + .5 \times (-3\ \mu(A|Q) + -1\ \mu(B|Q) + 0\ \mu(C|Q)) \geq .5, \end{aligned}$$

$$8\ \mu(A|R) + 4\ \mu(B|R) + 0\ \mu(C|R) \geq 0,$$

$$1\ \mu(A|Q) + 2\ \mu(B|Q) + 0\ \mu(C|Q) \geq 0.$$

These participational incentive constraints assert that the boss, the saleswoman with type R, and the saleswoman with type Q, must each get an expected payoff from the mediation plan that is not less than the best expected payoff that he or she could guarantee himself or herself without mediation. (The participational incentive constraints for the saleswoman are trivially redundant, since she controls no payoff-relevant actions, but we list them here for logical completeness.)

Subject to these participational incentive constraints, together with the informational incentive constraints and the probability constraints listed in section 3, the optimal mediation plan for the boss is as shown above, where he gets an expected payoff of 1.25 .

REFERENCES

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

R. Aumann [1987], "Correlated Equilibrium as an Expression of Bayesian Rationality," Econometrica 55, 1-18.

V. Crawford and J. Sobel [1982], "Strategic Information Transmission," Econometrica 50, 579-594.

J. Farrell [1986], "Meaning and Credibility in Cheap-Talk Games," GTE Laboratories working paper.

F. Forges [1986], "An Approach to Communication Equilibria," Econometrica 54, 1375-1385.

R. Myerson [1985], "Bayesian Equilibrium and Incentive-Compatibility: an Introduction," in L. Hurwicz, D. Schmeidler, and H. Sonnenschein, eds., Social Goals and Social Organization: Essays in Memory of Elisha Pazner, Cambridge University Press, pp. 229-259.

R. Myerson [1986], "Credible Negotiation Statements and Coherent Plans," Northwestern University working paper.