revelation principle

Communication is central to the economic problem (Hayek, 1945). Opportunities for mutually beneficial transactions cannot be found unless individuals share information about their preferences and endowments. Markets and other economic institutions should be understood as mechanisms for facilitating communication. However, people cannot be expected to reveal information when it is against their interests; for example, a seller may conceal his willingness to sell at a lower price. Rational behaviour in any specific communication mechanism can be analysed using game-theoretic equilibrium concepts, but efficient institutions can be identified only by comparison with all possible communication mechanisms. The revelation principle is a technical insight that allows us, in any given economic situation, to make general statements about all possible communication mechanisms.

The problem of making statements about all possible communication systems might seem intractably complex. Reports and messages may be expressed in rich languages with unbounded vocabulary. Communication systems can include both public announcements and private communication among smaller groups. Communication channels can have noise that randomly distorts messages. A communication mechanism may also specify how contractually enforceable transactions will depend on agents' reports and messages. So a general communication mechanism for any given set of agents may specify (*a*) a set of possible reports that each agent can send, (*b*) a set of possible messages that each agent can receive from the communication system, and (*c*) a probabilistic rule for determining the messages received and the enforceable transactions as a function of the reports sent by the agents. However, the revelation principle tells us that, for many economic purposes, it is sufficient for us to consider only a special class of mechanisms, called 'incentive-compatible direct-revelation mechanisms'.

In these mechanisms, every economic agent is assumed to communicate only with a central mediator. This mediator may be thought of as a trustworthy person or as a computer at the centre of a telephone network. In a direct-revelation mechanism, each individual is asked to report all of his private information confidentially to the mediator. After receiving these reports, the mediator then specifies all contractually enforceable transactions, as a function of these reports. If any individual controls private actions that are not contractually enforceable (such as efforts that others cannot observe), then the mediator also confidentially recommends an action to the individual. A direct-revelation mechanism is any rule for specifying how the mediator determines these contractual transactions and privately recommended actions, as a function of the private-information reports that the mediator receives.

A direct-revelation mechanism is said to be 'incentive compatible' if, when each individual expects that the others will be honest and obedient to the mediator, then no individual could ever expect to do better (given the information available to him) by reporting dishonestly to the mediator or by disobeying the mediator's recommendations. That is, the mechanism is incentive compatible if honesty and obedience is an equilibrium of the resulting communication game. The set of incentive-compatible direct-revelation mechanisms has good mathematical properties that often make it easy to analyse because it can be defined by a collection of linear inequalities, called 'incentive constraints'. Each of these incentive constraints expresses a requirement that an individual's expected utility from using a dishonest or disobedient strategy should not be greater than the individual's expected utility from being honest and obedient, when it is anticipated that everyone else will be honest and obedient.

The analysis of such incentive-compatible direct-revelation mechanisms might seem to be of limited interest, because real institutions rarely use such fully centralized mediation and often generate incentives for dishonesty or disobedience. For any equilibrium of any general communication mechanism, however, there exists an incentive-compatible direct-revelation mechanism that is essentially equivalent. This proposition is the revelation principle. Thus, the revelation principle tells us that, by analysing the set of incentive-compatible direct-revelation mechanisms, we can derive general properties of all equilibria of all coordination mechanisms.

The terms 'honesty' and 'obedience' here indicate two fundamental aspects of the general economic problem of communication. In a general communication system, an individual may send out messages or reports to share information that he knows privately, and he may also receive messages or recommendations to guide actions that he controls privately. The problem of motivating individuals to report their private information honestly is called 'adverse selection', and the problem of motivating individuals to implement their recommended actions obediently is called 'moral hazard'. To describe the intuition behind the revelation principle, let us consider first the special cases where only one or the other of these problems exists.

Pure adverse selection

First, let us formulate the revelation principle for the case of pure adverse selection, as developed in Bayesian social choice theory. In this case we are given a set of individuals, each of whom has some initial private information that may be called the individual's 'type', and there is a planning question of how a social allocation of resources should depend on the individuals' types. Each individual's payoff can depend on the resource allocation and on the types of all individuals according to some given utility function, and each type of each individual has some given probabilistic beliefs about the types of all other individuals. A general communication system would allow each individual i to send a message m_i in some rich language, and then the chosen resource allocation would depend on all these messages according to some rule $\gamma(m_1, \ldots, m_n)$. In any equilibrium of the game defined by this communication system, each individual i must have some strategy σ_i for choosing his message as a function of his type t_i, so that $m_i = \sigma_i(t_i)$.

For the given equilibrium $(\sigma_1, \ldots, \sigma_n)$ of the given social-choice rule γ, the revelation principle is satisfied by a mediation plan in which each individual is asked to confidentially report his type t_i to a central mediator, who then implements the social choice

$$\mu(t_1, \ldots, t_n) = \gamma(\sigma_1(t_1), \ldots, \sigma_n(t_n)).$$

So the mediator computes what message would be sent by the reported type of each individual i under his or her strategy σ_i, and then the mediator implements the resource allocation that would result from these messages under the rule γ. It is easy to see that honesty is an equilibrium under this mediation plan μ. If any individual could gain by lying to this mediator, when all others are expected to be honest, then this individual could have also gained by lying to himself when implementing his equilibrium strategy σ_i under the given mechanism γ, which would contradict the optimality condition that defines an equilibrium. So μ is an incentive-compatible direct-revelation mechanism that is equivalent to the given general mechanism γ with the given equilibrium $(\sigma_1, \ldots, \sigma_n)$.

In this case of pure adverse selection, the revelation principle was introduced by Gibbard (1973), but for a narrower solution concept (dominant strategies, instead of Bayesian equilibrium). The revelation principle for the broader solution concept of Bayesian equilibrium was recognized by Dasgupta, Hammond and Maskin (1979), Harris and Townsend (1981), Holmstrom (1977), Myerson (1979), and Rosenthal (1978).

Pure moral hazard

Next let us formulate the revelation principle for the case of pure moral hazard, as developed in Aumann's (1974) theory of correlated equilibrium. In this case we are given a set of individuals, each of whom controls some actions, and each individual's payoff can depend on the actions $(c_1, \ldots, c_n)$ that are chosen by all individuals, according to some given utility function $u_i(c_1, \ldots, c_n)$. That is, we are given a game in strategic form. In this case of pure moral hazard, nobody has any private information initially, but a communication process could give individuals different information before they choose their actions. In a general communication system, each individual i could get some message m_i in some rich language, with these messages $(m_1, \ldots, m_n)$ being randomly drawn from some joint probability distribution ρ. In any equilibrium of the game generated by adding this communication system, each individual i has some strategy σ_i for choosing his action c_i as a function of his message m_i, so that $c_i = \sigma_i(m_i)$.

For the given equilibrium $(\sigma_1, \ldots, \sigma_n)$ of the game with the given communication system ρ, the revelation principle is satisfied by a mediation plan in which the mediator randomly generates recommended actions in such a way that the probability of recommending actions $(c_1, \ldots, c_n)$ is the same as the probability of the given communication system ρ yielding messages $(m_1, \ldots, m_n)$ that would induce the players to choose $(c_1, \ldots, c_n)$ in the σ equilibrium. That is, the probability $\mu(c_1, \ldots, c_n)$ of the mediator recommending $(c_1, \ldots, c_n)$ is

$$\mu(c_1, \ldots, c_n) = \rho(\{(m_1, \ldots, m_n) | \sigma_1(m_1) = c_1, \ldots, \sigma_n(m_n) = c_n\}).$$

Then the mediator confidentially tells each individual i only which action c_i is recommended for him. Obedience is an equilibrium under this mediation plan μ because, if any individual could gain by disobeying this mediator when all others are expected to be obedient, then this individual could have also gained by disobeying himself in implementing his equilibrium strategy σ_i in the given game with communication system ρ. So μ is an incentive-compatible direct-revelation mechanism that is equivalent to the given mechanism ρ with the given equilibrium $(\sigma_1, \ldots, \sigma_n)$.

General formulations

Problems of adverse selection and moral hazard can be combined in the framework of Harsanyi's (1967) Bayesian games, where players have both types and actions. The revelation principle for general Bayesian games was formulated by Myerson (1982; 1985). A further generalization of the revelation principle to multistage games was formulated by Myerson (1986). In each case, the basic idea is that any equilibrium of any general communication system can be simulated by a maximally centralized communication system in which, at every stage, each individual confidentially reports all his private information to a central mediator, and then the mediator confidentially recommends an action to each individual, and the mediator's rule for generating recommendations from reports is designed so that honesty and obedience form an equilibrium of the mediated communication game.

The basic assumption here is that, although the motivations of all economic agents are problematic, we can find a mediator who is completely trustworthy and has no costs of processing information. Asking agents to reveal all relevant information to the trustworthy mediator maximizes the mediator's ability to implement any coordination plan. But telling any other agent more than is necessary to guide his choice of action would only increase the agent's ability to find ways of profitably deviating from the coordination plan.

For honesty and obedience to be an equilibrium, the mediation plan must satisfy incentive constraints which say that no individual could ever expect to gain by deviating to a strategy that involves lying to the mediator or disobeying a recommendation from the mediator. In a dynamic context, we must consider that an individual's most profitable deviation from honesty and obedience could be followed by further deviations in the future. So, to verify that an individual could never gain by lying, we must consider all possible deviation strategies in which the individual may thereafter choose actions that can depend disobediently on the mediator's recommendations (which may convey information about others' types and actions).

When we use sequential equilibrium as the solution concept for dynamic games with communication, the set of actions that can be recommended in a sequentially

incentive-compatible mechanism must be restricted somewhat. In a Bayesian game, if some action d_i could never be optimal for individual i to use when his type is t_i, no matter what information he obtained about others' types and actions, then obedience could not be sequentially rational in any mechanism where the mediator might ever recommend this action d_i to i after he reports type t_i. Myerson (1986) identified a larger set of *co-dominated actions* that can never be recommended in any sequentially incentive-compatible mechanism. Suppose that, if any individual observed a zero-probability event, then he could attribute this surprise to a mistake by the trembling hand of the mediator. Under this assumption, Myerson (1986) showed that the effect of requiring sequential rationality in games with communication is completely characterized by the requirement that no individuals should ever be expected to choose any co-dominated actions. (See Gerardi and Myerson, 2007.)

Limitations

The revelation principle says that each equilibrium of any communication mechanism is equivalent to the honest-obedient equilibrium of an incentive-compatible direct-revelation mechanism. But this direct-revelation mechanism may have other dishonest equilibria, which might not correspond to equilibria of the original mechanism. So the revelation principle cannot help us when we are concerned about the whole set of equilibria of a communication mechanism. Similarly, a given communication mechanism may have equilibria that change in some desirable way as we change the players' given beliefs about each others' types, but these different equilibria would correspond to different incentive-compatible mechanisms, and so this desirable property of the given mechanism could not be recognized with the revelation principle.

The assumption that perfectly trustworthy mediators are available is essential to the mathematical simplicity of the incentive-compatible set. Otherwise, if individuals can communicate only by making public statements that are immediately heard by everybody, then the set of equilibria may be smaller and harder to compute.

In principal-agent analysis we often apply the revelation principle to find the incentive-compatible mechanism that is optimal for the principal. If the principal would be tempted to use revealed information opportunistically, then there could be loss of generality in assuming that the agents reveal all their private information to the principal. But we should not confuse the principal with the mediator. The revelation principle can still be applied if the principal can get a trustworthy mediator to take the agents' reports and use them according to any specified mechanism.

There are often questions about whether the allocation selected by a mechanism could be modified by subsequent exchanges among the individuals. An individual's right to offer his possessions for sale at some future date could be accommodated in mechanism design by additional moral-hazard constraints.

For example, suppose the principal can sell an object each day, on days 1 and 2. The only buyer's value for such objects is either low \$1 or high \$3, low having probability 0.25. To maximize the principal's expected revenue with the buyer participating honestly, an optimal mechanism would sell both objects for \$3 if the buyer's type is high, but would sell neither if the buyer is low. But if no sale is recommended then the principal could infer that the buyer is low and would prefer to sell for \$1. Suppose now that the principal cannot be prevented from offering to sell for \$1 on either day. With these additional moral-hazard constraints, an optimal mechanism uses randomization by the mediator to conceal information from the principal. If the buyer reports low then the mediator recommends no sale on day 1 and selling for \$1 on day 2. If the buyer reports high, then with probability 1/3 the mediator recommends no sale on day 1 and selling for \$3 on day 2, but with probability 2/3 recommends selling for \$1.50 on both days. A no-sale recommendation on day 1 implies probability 0.5 of low, so that obedience yields the same expected revenue $0.5 \times (0 + 1) + 0.5 \times (0 + 3)$ as deviating to sell for \$1 on both days.

A proliferation of such moral-hazard constraints may greatly complicate the analysis, however. So in practice we often apply the revelation principle with an understanding that we may be overestimating the size of the feasible set, by assuming away some problems of mediator imperfection or moral hazard. When we use the revelation principle to show that a seemingly wasteful mechanism is actually efficient when incentive constraints are recognized, such overestimation of the incentive-feasible set would not weaken the impact of our results (as long as this mechanism remains feasible).

Centralized mediation is emphasized by the revelation principle as a convenient way of characterizing what people can achieve with communication, but this analytical convenience does not imply that centralization is necessarily the best way to coordinate an economy. For fundamental questions about socialist centralization versus free-market decentralization, we should be sceptical about an assumption that centralized control over national resources could not corrupt any mediator. The power of the revelation principle for such questions is instead its ability to provide a common analytical framework that applies equally to socialism and capitalism. For example, a standard result of revelation-principle analysis is that, if only one producer knows the production cost of a good, then efficient incentive-compatible mechanisms must allow this monopolistic producer to take positive informational rents or profits (Baron and Myerson, 1982). Thus the revelation principle can actually be used to support arguments for decentralized multi-source production, by showing that problems of profit-taking by an informational monopolist can be just as serious under socialism as under capitalism.

Nash (1951) advocated a different methodology for analysing communication in games. In Nash's approach, all opportunities for communication should be represented by moves in our extensive model of the dynamic game. Adding such communication moves may greatly increase the number of possible strategies for a player, because each strategy is a complete plan for choosing the player's moves throughout the game. But if all communication will occur in the implementation of these strategies, then the players' initial choices of their strategies must be independent. Thus, Nash argued, any dynamic

game can be normalized to a static strategic-form game, where players choose strategies simultaneously and independently, and Nash equilibrium is the general solution for such games.

With the revelation principle, however, communication opportunities are omitted from the game model and are instead taken into account by using incentive-compatible mechanisms as our solution concept. Characterizing the set of all incentive-compatible mechanisms is often easier than computing the Nash equilibria of a game with communication. Thus, by applying the revelation principle, we can get both a simpler model and a simpler solution concept for games with communication. But, when we use the revelation principle, strategic-form games are no longer sufficient for representing general dynamic games, because normalizing a game model to strategic form would suppress implicit opportunities for communicating during the game (see Myerson, 1986). So the revelation principle should be understood as a methodological alternative to Nash's strategic-form analysis.

ROGER B. MYERSON

See also **mechanism design.**

Bibliography

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

Baron, D. and Myerson, R. 1982. Regulating a monopolist with unknown costs. *Econometrica* 50, 911–30.

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

Gerardi, D. and Myerson, R. 2007. Sequential equilibrium in Bayesian games with communication. *Games and Economic Behavior* 60, 104–34.

Gibbard, A. 1973. Manipulation of voting schemes: a general result. *Econometrica* 41, 587–601.

Harris, M. and Townsend, R. 1981. Resource allocation under asymmetric information. *Econometrica* 49, 1477–99.

Harsanyi, J. 1967. Games with incomplete information played by Bayesian players. *Management Science* 14, 159–82, 320–34, 486–502.

Hayek, F. 1945. The use of knowledge in society. *American Economic Review* 35, 519–30.

Holmstrom, B. 1977. On incentives and control in organizations. Ph.D. thesis, Stanford University.

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

Myerson, R. 1982. Optimal coordination mechanisms in generalized principal-agent problems. *Journal of Mathematical Economics* 10, 67–81.

Myerson, R. 1985. Bayesian equilibrium and incentive compatibility: an introduction. In *Social Goals and Social Organization*, ed. L. Hurwicz, D. Schmeidler and H. Sonnenschein. Cambridge: Cambridge University Press.

Myerson, R. 1986. Multistage games with communication. *Econometrica* 54, 323–58.

Nash, J. 1951. Non-cooperative games. *Annals of Mathematics* 54, 286–95.

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

Shapley value

The *value* of an uncertain outcome (a 'lottery') is an a priori measure, in th participant's utility scale, of what he expects to obtain (this is the subject of 'utili theory'). The question is, how would one evaluate the prospects of a player in a mult person interaction, that is, in a game?

This question was originally addressed by Lloyd S. Shapley (1953a). The framewo was that of n-person games in coalitional form with side-payments, which are give by a set N of 'players', say 1, 2, ..., n, together with a 'coalitional function' v th associates to every subset S of N ('coalition') a real number $v(S)$, the maximal tot payoff the members of S can obtain (the 'worth' of S). An underlying assumption this model is that there exists a medium of exchange ('money') that is free transferable in unlimited amounts between the players, and moreover every player utility is additive with respect to it (that is, a transfer of x units from one player another decreases the first one's utility by x units and increases the second one's utili by x units; the total payoff of a coalition can thus be meaningfully defined as the su of the payoffs of its members). This requirement is known as existence of 'sid payments' or 'transferable utility'. In addition, the game is assumed to be adequate described by its coalitional function (that is, the worth $v(S)$ of each coalition S is we defined, and the abstraction from the extensive structure of the game to its coalition function leads to no essential loss; such a game is called a 'c-game'). The assumptions may be interpreted in a broader and more abstract sense. For example, a voting situation, a 'winning coalition' is assigned worth 1, and a 'losing' coalitio worth 0. The essential feature is that the prospects of each coalition may summarized by one number.

The *Shapley value* associates to each player in each such game a unique payoff – h 'value'. The value is required to satisfy the following four axioms. (EFF) *Efficiency Pareto optimality:* The sum of the values of all players equals $v(N)$, the worth of t grand coalition of all players (in a superadditive game $v(N)$ is the maximal amou that the players can jointly get); this axiom combines feasibility and efficiency. (SYM *Symmetry* or *equal treatment:* If two players in a game are substitutes (that is, t worth of no coalition changes when replacing one of the two players by the other on then their values are equal. (NUL) *Null* or *dummy player:* If a player in a game is su that the worth of every coalition remains the same when he joins it, then his value zero. (ADD) *Additivity:* The value of the sum of two games is the sum of the values the two games (equivalently, the value of a probabilistic combination of two games the same as the probabilistic combination of the values of the two games; this analogous to 'expected utility'). The surprising result of Shapley is that these fo axioms *uniquely determine* the values in *all* games.