Apaloo, J. 1997. Revisiting strategic models of evolution: the concept of neighborhood invader strategies. *Theoretical Population Biology* 52, 71–7.

Binmore, K., Samuelson, L. and Young, P. 2003. Equilibrium selection in bargaining models. *Games and Economic Behavior* 45, 296–328.

Bomze, I. 1995. Lotka–Volterra equation and replicator dynamics: new issues in classification. *Biological Cybernetics* 72, 447–53.

Bomze, I. and Pötscher, B. 1989. *Game Theoretical Foundations of Evolutionary Stability*. Lecture notes in economics and mathematical systems 324. Berlin: Springer-Verlag.

Cressman, R. 1992. *The Stability Concept of Evolutionary Games (A Dynamic Approach)*. Lecture notes in biomathematics 94. Berlin: Springer-Verlag.

Cressman, R. 2003. *Evolutionary Dynamics and Extensive Form Games*. Cambridge, MA: MIT Press.

Cressman, R. 2005. Continuously stable strategies, neighborhood superiority and two-player games with continuous strategy space. Mimeo.

Cressman, R. and Hofbauer, J. 2005. Measure dynamics on a one-dimensional continuous trait space: theoretical foundations for adaptive dynamics. *Theoretical Population Biology* 67, 47–59.

Dieckmann, U. and Law, R. 1996. The dynamical theory of coevolution: a derivation from stochastic ecological processes. *Journal of Mathematical Biology* 34, 579–612.

Eshel, I. 1983. Evolutionary and continuous stability. *Journal of Theoretical Biology* 103, 99–111.

Hofbauer, J., Schuster, P. and Sigmund, K. 1979. A note on evolutionarily stable strategies and game dynamics. *Journal of Theoretical Biology* 81, 609–12.

Hofbauer, J. and Sigmund, K. 1998. *Evolutionary Games and Population Dynamics*. Cambridge: Cambridge University Press.

Hofbauer, J. and Sigmund, K. 2003. Evolutionary game dynamics. *Bulletin of the American Mathematical Society* 40, 479–519.

Leimar, O. 2006. Multidimensional convergence stability and the canonical adaptive dynamics. In *Elements of Adaptive Dynamics*, ed. U. Dieckmann and J. Metz. Cambridge University Press.

Maynard Smith, J. 1982. *Evolution and the Theory of Games*. Cambridge: Cambridge University Press.

Maynard Smith, J. and Price, G. 1973. The logic of animal conflicts. *Nature* 246, 15–18.

Morris, S., Rob, R. and Shin, H. 1995. Dominance and belief potential. *Econometrica* 63, 145–57.

Nachbar, J. 1992. Evolution in the finitely repeated Prisoner's Dilemma. *Journal of Economic Behavior and Organization* 19, 307–26.

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

Oechssler, J. and Riedel, F. 2002. On the dynamic foundation of evolutionary stability in continuous models. *Journal of Economic Theory* 107, 223–52.

Sandholm, W. 2006. *Population Games and Evolutionary Dynamics*. Cambridge, MA: MIT Press.

Selten, R. 1980. A note on evolutionarily stable strategies in asymmetrical animal contests. *Journal of Theoretical Biology* 84, 93–101.

Selten, R. 1983. Evolutionary stability in extensive two-person games. *Mathematical Social Sciences* 5, 269–363.

Taylor, P. and Jonker, L. 1978. Evolutionarily stable strategies and game dynamics. *Mathematical Biosciences* 40, 145–156.

Thomas, B. 1985. On evolutionarily stable sets. *Journal of Mathematical Biology* 22, 105–15.

van Damme, E. 1991. *Stability and Perfection of Nash Equilibria*, 2nd edn. Berlin: Springer-Verlag.

Weibull, J. 1995. *Evolutionary Game Theory*. Cambridge, MA: MIT Press.

Zeeman, E. 1980. Population dynamics from game theory. In *Global Theory of Dynamical Systems*, eds. Z. Nitecki and C. Robinson. Lecture notes in mathematics 819. Berlin: Springer.

# mechanism design

## Overview

A mechanism is a specification of how economic decisions are determined as a function of the information that is known by the individuals in the economy. In this sense, almost any kind of market institution or economic organization can be viewed, in principle, as a mechanism. Thus mechanism theory can offer a unifying conceptual structure in which a wide range of institutions can be compared, and optimal institutions can be identified.

The basic insight of mechanism theory is that *incentive constraints* should be considered coequally with *resource constraints* in the formulation of the economic problem. In situations where individuals' private information and actions are difficult to monitor, the need to give people an incentive to share information and exert efforts may impose constraints on economic systems just as much as the limited availability of raw materials. The theory of mechanism design is the fundamental mathematical methodology for analysing these constraints.

The study of mechanisms begins with a special class of mechanisms called *direct-revelation* mechanisms, which operate as follows. There is assumed to be a mediator who can communicate separately and confidentially with every individual in the economy. This mediator may be thought of as a trustworthy person, or as a computer tied into a telephone network. At each stage of the economic process, each individual is asked to report all of his private information (that is, everything that he knows that other individuals in the economy might not know) to the mediator. After receiving these reports confidentially from every individual, the mediator may then confidentially recommend some action or move to each individual. A direct-revelation mechanism is any rule for specifying how the mediator's recommendations are determined, as a function of the reports received.

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, if honesty and obedience is an equilibrium (in the game-theoretic sense), then the mechanism is incentive compatible.

The analysis of such incentive-compatible direct-revelation mechanisms might at first seem to be of rather narrow interest, because such fully centralized mediation of economic systems is rare, and incentives for dishonesty and disobedience are commonly observed in real economic institutions. The importance of studying such mechanisms is derived from two key insights: (i) for any equilibrium of any general mechanism, there is an incentive-compatible direct-revelation mechanism that is

essentially equivalent; and (ii) the set of incentive-compatible direct-revelation mechanisms has simple mathematical properties that often make it easy to characterize, because it can be defined by a set of linear inequalities. Thus, by analysing incentive-compatible direct-revelation mechanisms, we can characterize what can be accomplished in all possible equilibria of all possible mechanisms, for a given economic situation.

Insight (i) above is known as the *revelation principle*. It was first recognized by Gibbard (1973), but for a somewhat narrower solution concept (dominant strategies, instead of Bayesian equilibrium) and for the case where only informational honesty is problematic (no moral hazard). The formulation of the revelation principle for the broader solution concept of Bayesian equilibrium, but still in the case of purely informational problems, was recognized independently by many authors around 1978 (see Dasgupta, Hammond and Maskin, 1979; Harris and Townsend, 1981; Holmstrom, 1977; Myerson, 1979; Rosenthal, 1978). Aumann's (1974; 1987) concept of *correlated equilibrium* gave the first expression to the revelation principle in the case where only obedient choice of actions is problematic (pure moral hazard, no adverse selection). The synthesis of the revelation principle for general Bayesian games with incomplete information, where both honesty and obedience are problematic, was given by Myerson (1982). A generalization of the revelation principle to multistage games was stated by Myerson (1986).

The intuition behind the revelation principle is as follows. First, a central mediator who has collected all relevant information known by all individuals in the economy could issue recommendations to the individuals so as to simulate the outcome of any organizational or market system, centralized or decentralized. After the individuals have revealed all of their information to the mediator, he can simply tell them to do whatever they would have done in the other system. Second, the more information that an individual has, the harder it may be to prevent him from finding ways to gain by disobeying the mediator. So the incentive constraints will be least binding when the mediator reveals to each individual only the minimal information needed to identify his own recommended action, and nothing else about the reports or recommendations of other individuals. So, if we assume that the mediator is a discrete and trustworthy information-processing device, with no costs of processing information, then there is no loss of generality in assuming that each individual will confidentially reveal all of his information to the mediator (maximal revelation to the trustworthy mediator), and the mediator in return will reveal to each individual only his own recommended action (minimal revelation to the individuals whose behaviour is subject to incentive constraints).

The formal proof of the revelation principle is difficult only because it is cumbersome to develop the notation for defining, in full generality, the set of all general mechanisms, and for defining equilibrium behaviour by the individuals in any given mechanism. Once all of this notation is in place, the construction of the equivalent incentive-compatible direct-revelation mechanism is straightforward. Given any mechanism and any equilibrium of the mechanism, we simply specify



that the mediator's recommended actions are those that would result in the given mechanism if everyone behaved as specified in the given equilibrium when his actual private information was as reported to the mediator. To check that this constructed direct-revelation mechanism is incentive compatible, notice that any player who could gain by disobeying the mediator could also gain by similarly disobeying his own strategy in the given equilibrium of the given mechanism, which is impossible (by definition of equilibrium).

## Mathematical formulations

Let us offer a precise general formulation of the proof of the revelation principle in the case where individuals have private information about which they could lie, but there is no question of disobedience of recommended actions or choices. For a general model, suppose that there are $n$ individuals, numbered 1 to $n$. Let $C$ denote the set of all possible combinations of actions or resource allocations that the individuals may choose in the economy. Each individual in the economy may have some private information about his preferences and endowments, and about his beliefs about other individuals' private information. Following Harsanyi (1967), we may refer to the state of an individual's private information as his *type*. Let $T_i$ denote the set of possible types for any individual $i$, and let $T = T_1 \times \cdots \times T_n$ denote the set of all possible combinations of types for all individuals.

The preferences of each individual $i$ may be generally described by some *payoff function* $u_i : C \times T \to \mathbb{R}$, where $u_i(c,(t_1,\ldots, t_n))$ denotes the payoff, measured in some von Neumann–Morgenstern utility scale, that individual $i$ would get if $c$ was the realized resource allocation in $C$ when $(t_1,\ldots, t_n)$ denotes the actual types of the individuals $1,\ldots, n$ respectively. For short, we may write $t = (t_1,\ldots, t_n)$ to describe a combination of types for all individuals.

The beliefs of each individual $i$, as a function of his type, may be generally described by some function $p_i(\cdot\,|\,\cdot)$, where $p_i(t_1,\ldots, t_{i-1}, t_{i+1},\ldots, t_n|t_i)$ denotes the probability that individual $i$ would assign to the event that the other individuals have types as in $(t_1,\ldots, t_{i-1}, t_{i+1},\ldots, t_n)$, when $i$ knows that his own type is $t_i$. For short, we may write $t_{-i} = (t_1,\ldots, t_{i-1}, t_{i+1},\ldots, t_n)$, to describe a combination of types for all individuals other than $i$. We may let $T_{-i} = T_1 \times \cdots \times T_{i-1} \times T_{i+1} \times \cdots \times T_n$ denote the set of all possible combinations of types for the individuals other than $i$.

The general model of an economy defined by these structures $(C, T_1,\ldots, T_n, u_1,\ldots, u_n, p_1,\ldots, p_n)$ is called a Bayesian collective-choice problem.

Given a Bayesian collective-choice problem, a general mechanism would be any function of the form $\gamma$: $S_1 \times \cdots \times S_n \to C$, where, for each $i$, $S_i$ is a nonempty set that denotes the set of strategies that are available for individual $i$ in this mechanism. That is, a general mechanism specifies the strategic options that each individual may choose among, and the social choice or allocation of resources that would result from any combination of strategies that the individuals might choose. Given a mechanism, an equilibrium is any specification of how each individual may choose his strategy in the

mechanism as a function of his type, so that no individual, given only his own information, could expect to do better by unilaterally deviating from the equilibrium. That is, $\sigma = (\sigma_1, \ldots, \sigma_n)$ is an equilibrium of the mechanism $\gamma$ if, for each individual $i$, $\sigma_i$ is a function from $T_i$ to $S_i$, and, for every $t_i$ in $T_i$ and every $s_i$ in $S_i$,

$$\Sigma_{t_{-i} \in T_{-i}} \; p_i(t_{-i}|t_i) u_i(\gamma(\sigma(t)), t) \geq \Sigma_{t_{-i} \in T_{-i}} \; p_i(t_{-i}|t_i) u_i(\gamma(\sigma_{-i}(t_{-i}), s_i), t).$$

(Here $\sigma(t) = (\sigma_1(t_1), \ldots, \sigma_n(t_n))$ and $(\sigma_{-i}(t_{-i}), s_i) = (\sigma_1(t_1), \ldots, \sigma_{i-1}(t_{i-1}), s_i, \sigma_{i+1}(t_{i+1}), \ldots, \sigma_n(t_n))$.) Thus, in an equilibrium $\sigma$, no individual $i$, knowing only his own type $t_i$, could increase his expected payoff by changing his strategy from $\sigma_i(t_i)$ to some other strategy $s_i$, when he expects all other individuals to behave as specified by the equilibrium $\sigma$. (This concept of equilibrium is sometimes often called *Bayesian equilibrium* because it respects the assumption that each player knows only his own type when he chooses his strategy in $S_i$. For a comparison with other concepts of equilibrium, see Dasgupta, Hammond and Maskin, 1979, and Palfrey and Srivastava, 1987).

In this context, a direct-revelation mechanism is any mechanism such that the set $S_i$ of possible strategies for each player $i$ is the same as his set of possible types $T_i$. A direct-revelation mechanism is (Bayesian) incentive-compatible iff it is an equilibrium (in the Bayesian sense defined above) for every individual always to report his true type. Thus, $\mu$: $T_1 \times \ldots \times T_n \rightarrow C$ is an incentive-compatible direct-revelation mechanism if, for each individual $i$ and every pair of types $t_i$ and $r_i$ in $T_i$,

$$\Sigma_{t_{-i} \in T_{-i}} \; p_i(t_{-i}|t_i) u_i(\mu(t), t) \geq \Sigma_{t_{-i} \in T_{-i}} \; p_i(t_{-i}|t_i) u_i(\mu(t_{-i}, r_i), t).$$

(Here $(t_{-i}, r_i) = (t_1, \ldots, t_{i-1}, r_i, t_{i+1}, \ldots, t_n)$.) We may refer to these constraints as the *informational incentive constraints* on the direct-revelation mechanism $\mu$. These informational incentive constraints are the formal representation of the economic problem of *adverse selection*, so they may also be called adverse-selection constraints (or self-selection constraints).

Now, to prove the revelation principle, given any general mechanism $\gamma$ and any Bayesian equilibrium $\sigma$ of the mechanism $\gamma$, let $\mu$ be the direct-revelation mechanism $\mu$ defined so that, for every $t$ in $T$,

$$\mu(t) = \gamma(\sigma(t)).$$

Then this mechanism $\mu$ always leads to the same social choice as $\gamma$ does, when the individuals behave as in the equilibrium $\sigma$. Furthermore, $\mu$ is incentive compatible because, for any individual $i$ and any two types $t_i$ and $r_i$ in $T_i$,

$$\begin{aligned}
\Sigma_{t_{-i} \in T_{-i}} \; p_i(t_{-i}|t_i) u_i(\mu(t), t) &= \Sigma_{t_{-i} \in T_{-i}} \; p_i(t_{-i}|t_i) u_i(\gamma(\sigma(t)), t) \\
&\geq \Sigma_{t_{-i} \in T_{-i}} \; p_i(t_{-i}|t_i) u_i(\gamma(\sigma_{-i}(t_{-i}), \sigma_i(r_i)), t) \\
&= \Sigma_{t_{-i} \in T_{-i}} \; p_i(t_{-i}|t_i) u_i(\mu(t_{-i}, r_i), t).
\end{aligned}$$

Thus, $\mu$ is an incentive-compatible direct-revelation mechanism that is equivalent to the given mechanism $\gamma$ with its equilibrium $\sigma$.

Notice that the revelation principle asserts that any pair consisting of a mechanism *and* an equilibrium is equivalent to an incentive-compatible direct-revelation mechanism. Thus, a general mechanism that has several equilibria may correspond to several different incentive-compatible mechanisms, depending on which equilibrium is considered.

Furthermore, the same general mechanism will generally have different equilibria in the context of different Bayesian collective-choice problems, where the structure of individuals' beliefs and payoffs are different. For example, consider a first-price sealed-bid auction where there are five potential bidders who are risk-neutral with independent private values drawn from the same distribution over \$0 to \$10. If the bidders' values are drawn from a uniform distribution over this interval, then there is an equilibrium in which each bidder bids 4/5 of his value. On the other hand, if the bidders' values are drawn instead from a distribution with a probability density that is proportional to the square of the value, then there is an equilibrium in which each bidder bids 8/9 of his value. So in one situation the first-price sealed-bid auction (a general mechanism) corresponds to an incentive-compatible mechanism in which the bidder who reports the highest value gets the object for 4/5 of his reported value; but in the other situation it corresponds to an incentive-compatible mechanism in which the bidder who reports the highest value gets the object for 8/9 of his reported value. There is no incentive-compatible direct-revelation mechanism that is equivalent to the first-price sealed-bid auction in all situations, independently of the bidders' beliefs about each others' values. Thus, if we want to design a mechanism that has good properties in the context of many different Bayesian collective-choice problems, we cannot necessarily restrict our attention to incentive-compatible direct-revelation mechanisms, and so our task is correspondingly more difficult. (See Wilson, 1985, for a remarkable effort at this kind of difficult question.)

Even an incentive-compatible mechanism itself may have other dishonest equilibria that correspond to different incentive-compatible mechanisms. Thus, when we talk about selecting an incentive-compatible mechanism and assume that it will then be played according to its honest equilibrium, we are implicitly making an assumption about the selection of an equilibrium as well as of a mechanism or communication structure. Thus, for example, when we say that a particular incentive-compatible mechanism maximizes a given individual's expected utility, we mean that, if you could choose any general mechanism for coordinating the individuals in the economy and if you could also (by some public statement, as a focal arbitrator, using Schelling's, 1960, *focal-point effect*) designate the equilibrium that the individuals would play in your mechanism, then you could not give this given individual a higher expected utility than by choosing this incentive-compatible mechanism and its honest equilibrium.

In many situations, an individual may have a right to refuse to participate in an economic system or organization. For example, a consumer generally has the right to refuse to participate in any trading scheme and instead just consume his initial

endowment. If we let $w_i(t_i)$ denote the utility payoff that individual $i$ would get if he refused to participate when his type is $t_i$, and if we assume that an individual can make the choice not to participate after learning his type, then an incentive-compatible mechanism $\mu$ must also satisfy the following constraint, for every individual $i$ and every possible type $t_i$:

$$\Sigma_{t_{-i}\in T_{-i}}\ p_i(t_{-i}|t_i)u_i(\mu(t),t) \geq w_i(t_i).$$

These constraints are called *participational incentive constraints*, or *individual-rationality* constraints.

In the analysis of Bayesian collective-choice problems, we have supposed that the only incentive problem was to get people to share their information, and to agree to participate in the mechanism in the first place. More generally, a social choice may be privately controlled by one or more individuals who cannot be trusted to follow some pre-specified plan when it is not in their best interests. For example, suppose now that the choice in $C$ is privately controlled by some individual (call him 'individual 0') whose choice of an action in $C$ cannot be regulated. To simplify matters here, let us suppose that this individual 0 has no private information. Let $p_0(t)$ denote the probability that this individual would assign to the event that $t=(t_1,\ldots,t_n)$ is the profile of types for the other $n$ individuals, and let $u_0(c, t)$ denote the utility payoff that this individual receives if he chooses action $c$ when $t$ is the actual profile of types. Then, to give this active individual an incentive to obey the recommendations of a mediator who is implementing the direct-revelation mechanism $\mu$, $\mu$ must satisfy

$$\Sigma_{t\in T}\ p_0(t)u_0(\mu(t),t) \geq \Sigma_{t\in T}\ p_0(t)u_0(\delta(\mu(t)),t)$$

for every function $\delta$: $C\rightarrow C$. These constraints assert that obeying the actions recommended by the mediator is better for this individual than any disobedient strategy $\delta$ under which he would choose $\delta(c)$ if the mediator recommended $c$. Such constraints are called *strategic incentive constraints* or *moral-hazard constraints*, because they are the formal representation of the economic problem of moral hazard.

For a formulation of general incentive constraints that apply when individuals both have private information and control private actions, see Myerson (1982) or (1985).

## Applications

In general, the mechanism-theoretic approach to economic problems is to list the constraints that an incentive-compatible mechanism must satisfy, and to try to characterize the incentive-compatible mechanisms that have properties of interest.

For example, one early contribution of mechanism theory was the derivation of general *revenue equivalence* theorems in auction theory. Ortega-Reichert (1968) found that, when bidders are risk-neutral and have private values for the object being sold that are independent and drawn from the same distribution, then a remarkably diverse collection of different auction mechanisms all generate the same expected revenue to the seller, when bidders use equilibrium strategies. In all of these different mechanisms and equilibria, it turned out that the bidder whose value for the object was highest would always end up getting the object, while a bidder whose value for the object was zero would never pay anything. By analysing the incentive constraints, Harris and Raviv (1981), Myerson (1981) and Riley and Samuelson (1981) showed that all incentive-compatible mechanisms with these properties would necessarily generate the same expected revenue, in such economic situations.

Using methods of constrained optimization, the problem of finding the incentive-compatible mechanism that maximizes some given objective (one individual's expected utility, or some social welfare function) can be solved for many examples. The resulting optimal mechanisms often have remarkable qualitative properties.

For example, suppose a seller, with a single indivisible object to sell, faces five potential buyers or bidders, whose private values for the object are independently drawn from a uniform distribution over the interval from \$0 to \$10. If the objective is to maximize the sellers' expected revenue, optimal auction mechanisms exist and all have the property that the object is sold to the bidder with the highest value for it, except that the seller keeps the object in the event that the bidders' values are all less than \$5. Such a result may seem surprising, because this event could occur with positive probability (1/32) and in this event the seller is getting no revenue in an 'optimal' auction, even though any bidder would almost surely be willing to pay him a positive price for the object. Nevertheless, no incentive-compatible mechanism (satisfying the participational and informational incentive constraints) can offer the seller higher expected utility than these optimal auctions, and thus no equilibrium of any general auction mechanism can offer higher expected revenue either. Maximizing expected revenue requires a positive probability of seemingly wasteful allocation.

The threat of keeping the object, when all bidders report values below \$5, increases the seller's expected revenue because it gives the bidders an incentive to bid higher and pay more when their values are above \$5. In many other economic environments, we can similarly prove the optimality of mechanisms in which seemingly wasteful threats are carried out with positive probability. People have intuitively understood that costly threats are often made to give some individual an incentive to reveal some information or choose some action, and the analysis of incentive constraints allows us to formalize this understanding rigorously.

In some situations, incentive constraints imply that such seemingly wasteful allocations may have to occur with positive probability in all incentive-compatible mechanisms, and so also in all equilibria of all general mechanisms. For example, Myerson and Satterthwaite (1983) considered bilateral bargaining problems between a seller of some object and a potential buyer, both of whom are risk-neutral and have independent private values for the object that are drawn out of distributions that have continuous positive probability densities over some pair of intervals that have an intersection of positive length. Under these technical (but apparently quite weak) assumptions, it is impossible to satisfy the participational and informational incentive constraints with any mechanism in which the buyer gets the object whenever it is worth more to him than to the seller. Thus, we cannot hope to guarantee the attainment of full *ex post* efficiency of resource allocations in bilateral bargaining

problems where the buyer and seller are uncertain about each other's reservation prices. If we are concerned with welfare and efficiency questions, it may be more productive to try to characterize the incentive-compatible mechanisms that maximize the expected total gains from trade, or that maximize the probability that a mutually beneficial trade will occur. For example, in the bilateral bargaining problem where the seller's and buyer's private values for the object are independent random variables drawn from a uniform distribution over the interval from \$0 to \$10, both of these objectives are maximized subject to incentive constraints by mechanisms in which the buyer gets the object if and only if his value is greater than the seller's value by \$2.50 or more. Under such a mechanism, the event that the seller will keep the object when it is actually worth more to the buyer has probability 7/32, but no equilibrium of any general mechanism can generate a lower probability of this event.

The theory of mechanism design has fundamental implications about the domain of applicability of Coase's (1960) theorem, which asserts the irrelevance of initial property rights to efficiency of final allocations. The unavoidable possibility of failure to realize mutually beneficial trades, in such bilateral trading problems with two-sided uncertainty, can be interpreted as one of the 'transaction costs' that limits the validity of Coase's theorem. Indeed, as Samuelson (1985) has emphasized, reassignment of property rights generally changes the payoffs that individuals can guarantee themselves without selling anything, which changes the right-hand sides of the participational incentive constraints, which in turn can change the maximal social welfare achievable by an optimal incentive-compatible mechanism.

For example, consider again the case where there is one object and two individuals who have private values for the object that are independent random variables drawn from a uniform distribution over the interval from \$0 to \$10. When we assumed above that one was the 'seller', we meant that he had the right to keep the object and pay nothing to anyone, until he agreed to some other arrangement. Now, let us suppose instead that the rights to the object are distributed equally between the two individuals. Suppose that the object is a divisible good and each individual has a right to take half of the good and pay nothing, unless he agrees to some other arrangement. (Assume that, if an individual's value for the whole good is $t_i$, then his value for half would be $t_i/2$.) With this symmetric assignment of property rights, we can design incentive-compatible mechanisms in which the object always ends up being owned entirely by the individual who has the higher value for it, as Cramton, Gibbons and Klemperer (1987) have shown.

For example, consider the game in which each individual independently puts money in an envelope, and then the individual who put more money in his envelope gets the object, while the other individual takes the money in both envelopes. This game has an equilibrium in which each individual puts into his envelope an amount equal to one-third of his value for the whole good. This equilibrium of this game is equivalent to an incentive-compatible direct-revelation mechanism in which the individual who reports the higher value pays one-third of his value to buy out the other individual's half-share. This mechanism would violate the participational incentive constraints if one individual had a right to the whole good (in which case, for example, if his value were \$10 then he would be paying \$3.33 under this mechanism for a good that he already owned). But with rights to only half of the good, no type of either individual could expect to do better (at the beginning of the game, when he knows his own value but not the other's) by keeping his half and refusing to participate in this mechanism.

More generally, redistribution of property rights tends to reduce the welfare losses caused by incentive constraints when it creates what Lewis and Sappington (1989) have called *countervailing incentives*. In games where one individual is the seller and the other is the buyer, if either individual has an incentive to lie, it is usually because the seller wants to overstate his value or the buyer wants to understate his value. In the case where either individual may buy the other's half-share, neither individual can be sure at first whether he will be the buyer or the seller (unless he has the highest or lowest possible value). Thus, a buyer-like incentive to understate values, in the event where the other's value is lower, may help to cancel out a seller-like incentive to overstate values, in the event where the other's value is higher.

The theory of mechanism design can also help us to appreciate the importance of mediation in economic relationships and transactions. There are situations in which, if the individuals were required to communicate with each other only through perfect noiseless communication channels (for example, in face-to-face dialogue), then the set of all possible equilibria would be much smaller than the set of incentive-compatible mechanisms that are achievable with a mediator. (Of course, the revelation principle asserts that the former set cannot be larger than the latter.)

For example, consider the following 'sender–receiver game' due to J. Farrell. Player 1 has a privately known type that may be $\alpha$ or $\beta$, but he has no payoff-relevant action to choose. Player 2 has no private information, but he must choose an action from the set $\{x, y, z\}$. The payoffs to players 1 and 2 respectively depend on 1's type and 2's action as follows.

| | $x$ | $y$ | $z$ |
|---|---|---|---|
| $\alpha$ | 2, 3 | 1, 2 | 0, 0 |
| $\beta$ | 4, −3 | 8, −1 | 0, 0 |

At the beginning of the game, player 2 believes that each of 1's two possible types has probability 1/2.

Suppose that, knowing his type, player 1 is allowed to choose a message in some arbitrarily rich language, and player 2 will hear player 1's message (with no noise or distortion) before choosing his action. In every equilibrium of this game, including the randomized equilibria, player 2 must choose $y$ with probability 1, after every message that player 1 may choose in equilibrium (see Farrell, 1993; Myerson, 1988). If there were some message that player 1 could use to increase the probability of player 2 choosing $x$ (for example, 'I am $\alpha$, so choosing $x$ would be best for us both!'), then he would always send such a message when his type was $\alpha$. (It can be shown that no

message could ever induce player 2 to randomize between $x$ and $z$.) So not receiving such a message would lead 2 to infer that 1's type was $\beta$, which implies that 2 would rationally choose $z$ whenever such a message was not sent, so that both types of 1 should always send the message (any randomization between $x$ and $y$ is better than $z$ for both types of 1). But a message that is always sent by player 1, no matter what his type is, would convey no information to player 2, so that 2 would rationally choose his *ex ante* optimal action $y$.

If we now allow the players to communicate through a mediator who uses a randomized mechanism, then we can apply the revelation principle to characterize the surprisingly large set of possible incentive-compatible mechanisms. Among all direct-revelation mechanisms that satisfy the relevant informational incentive constraints for player 1 and strategic incentive constraints for player 2, the best for player 2 is as follows: if player 1 reports to the mediator that his type is $\alpha$ then with probability 2/3 the mediator recommends $x$ to player 2, and with probability 1/3 the mediator recommends $y$ to player 2; if player 1 reports to the mediator that his type is $\beta$ then with probability 2/3 the mediator recommends $y$ to player 2, and with probability 1/3 the mediator recommends $z$ to player 2. Notice that this mechanism is also better for player 1 than the unmediated equilibria when 1's type is $\alpha$, although it is worse for 1 when his type is $\beta$.

Other mechanisms that player 2 might prefer would violate the strategic incentive constraint that player 2 should not expect to gain by choosing $z$ instead of $y$ when $y$ is recommended. If player 2 could pre-commit himself always to obey the mediator's recommendations, then better mechanisms could be designed.

## Efficiency

The concept of efficiency becomes more difficult to define in economic situations where individuals have different private information at the time when the basic decisions about production and allocation are made. A welfare economist or social planner who analyses the Pareto efficiency of an economic system must use the perspective of an outsider, so he cannot base his analysis on the individuals' private information. Otherwise, public testimony as to whether an economic mechanism or its outcome would be 'efficient' could implicitly reveal some individuals' private information to other individuals, which could in turn alter their rational behaviour and change the outcome of the mechanism! Thus, Holmstrom and Myerson (1983) argued that efficiency should be considered as a property of mechanisms, rather than of the outcome or allocation ultimately realized by the mechanism (which will depend on the individuals' private information).

Thus, a definition of Pareto efficiency in a Bayesian collective-choice problem must look something like this: 'a mechanism is efficient if there is no other feasible mechanism that may make some other individuals better off and will certainly not make other individuals worse off.' However, this definition is ambiguous in at least two ways.



First, we must specify whether the concept of feasibility takes incentive constraints into account or not. The concept of feasibility that ignores incentive constraints may be called *classical feasibility*. In these terms, the fundamental insight of mechanism theory is that incentive constraints are just as real as resource constraints, so that incentive compatibility may be a more fruitful concept than classical feasibility for welfare economics.

Second, we must specify what information is to be considered in determining whether an individual is 'better off' or 'worse off'. One possibility is to say that an individual is made worse off by a change that decreases his expected utility payoff as would be computed before his own type or any other individuals' types are specified. This is called the *ex ante* welfare criterion. A second possibility is to say that an individual is made worse off by a change that decreases his conditionally expected utility, given his own type (but not given the types of any other individuals). An outside observer, who does not know any individual's type, would then say that an individual may be made worse off, in this sense, if this conditionally expected utility were decreased for at least one possible type of the individual. This is called the *interim* welfare criterion. A third possibility is to say that an individual is made worse off by a change that decreases his conditionally expected utility given the types of all individuals. An outside observer would then say that an individual may be worse off in this sense if his conditionally expected utility were decreased for at least one possible combination of types for all the individuals. This is called the *ex post* welfare criterion.

If each individual knows his own type at the time when economic plans and decisions are made, then the interim welfare criterion should be most relevant to a social planner. Thus, Holmstrom and Myerson (1983) argue that, for welfare analysis in a Bayesian collective-choice problem, the most appropriate concept of efficiency is that which combines the interim welfare criterion and the incentive-compatible definition of feasibility. This concept is called *incentive efficiency*, or *interim incentive efficiency*. That is, a mechanism $\mu$: $T \to C$ is incentive efficient if it is an incentive-compatible mechanism and there does not exist any other incentive-compatible mechanism $\gamma$: $T \to C$ such that for every individual $i$ and every type $t_i$ in $T_i$,

$$\Sigma_{t_{-i} \in T_{-i}} \; p_i(t_{-i}|t_i) u_i(\gamma(t), t) \geq \Sigma_{t_{-i} \in T_{-i}} \; p_i(t_{-i}|t_i) u_i(\mu(t), t),$$

and there is at least one type of at least one individual for which this inequality is strict. If a mechanism is incentive efficient, then it cannot be common knowledge among the individuals, at the stage when each knows only his own type, that there is some other incentive-compatible mechanism that no one would consider worse (given his own information) and some might consider strictly better.

For comparison, another important concept is classical *ex post* efficiency, defined using the *ex post* welfare criterion and the classical feasibility concept. That is, a mechanism $\mu$: $T \to C$ is (*classically*) *ex post efficient* iff there does not exist any other mechanism $\gamma$: $T \to C$ (not necessarily incentive compatible) such that, for every individual $i$ and every combination of individuals' types $t$ in $T = T_1 \times \cdots \times T_n$,

$$u_i(\gamma(t), t) \geq u_i(\mu(t), t),$$

with strict inequality for at least one individual and at least one combination of individuals' types.

The appeal of *ex post* efficiency is that there may seem to be something unstable about a mechanism that sometimes leads to outcomes such that, if everyone could share their information, they could identify another outcome that would make them all better off. However, we have seen that bargaining situations exist where no incentive-compatible mechanisms are *ex post* efficient. In such situations, the incentive constraints imply that rational individuals would be unable to share their information to achieve these gains, because if everyone were expected to do so then at least one type of one individual would have an incentive to lie.

Thus, a benevolent outside social planner who is persuaded by the usual Paretian arguments should choose some incentive-efficient mechanism. To determine more specifically an 'optimal' mechanism within this set, a social welfare function is needed that defines tradeoffs, not only between the expected payoffs of different individuals but also between the expected payoffs of different types of each individual. That is, given any positive utility-weights $\lambda_i(t_i)$ for each type $t_i$ of each individual $i$, one can generate an incentive-efficient mechanism by maximizing

$$\Sigma_{i=1}^{n} \; \Sigma_{t_i \in T_i} \; \lambda_i(t_i) \; \Sigma_{t_{-i} \in T_{-i}} \; p_i(t_{-i}|t_i) u_i(\mu(t), t)$$

over all $\mu$: $T \rightarrow C$ that satisfy the incentive constraints; but different vectors of utility weights may generate different incentive-efficient mechanisms.

### Bargaining over mechanisms

A positive economic theory must go beyond welfare economics and try to predict the economic institutions that may actually be chosen by the individuals in an economy. Having established that a social planner can restrict his attention to incentive-compatible direct-revelation mechanisms, which is a mathematically simple set, it is natural to assume that rational economic agents who are themselves negotiating the structure of their economic institutions should be able to bargain over the set of incentive-compatible direct-revelation mechanisms. But if we assume that individuals know their types already at the time when fundamental economic plans and decisions are made, then we need a theory of mechanism selection by individuals who have private information.

When we consider bargaining games in which individuals can bargain over mechanisms, there should be no loss of generality in restricting our attention to equilibria in which there is one incentive-compatible mechanism that is selected with probability 1 independently of anyone's type. This proposition, called the *inscrutability principle*, can be justified by viewing the mechanism-selection process as itself part of a more broadly defined general mechanism and applying the revelation principle. For example, suppose that there is an equilibrium of the mechanism-selection game in which some mechanism $\mu$ would be chosen if individual 1's type were $\alpha$ and some other mechanism $\nu$ would be chosen if 1's type were $\beta$. Then there should exist an equivalent equilibrium of the mechanism-selection game in which the individuals always select a direct-revelation mechanism that coincides with mechanism $\mu$ when individual 1 confidentially reports type $\alpha$ to the mediator (in the implementation of the mechanism, after it has been selected), and that coincides with mechanism $\nu$ when 1 reports type $\beta$ to the mediator.

However, the inscrutability principle does not imply that the possibility of revealing information during a mechanism-selection process is irrelevant. There may be some mechanisms that we should expect not to be selected by the individuals in such a process, precisely because some individuals would choose to reveal information about their types rather than let these mechanisms be selected. For example, consider the following Bayesian collective-choice problem, due to Holmstrom and Myerson (1983). There are two individuals, 1 and 2, each of whom has two possible types, $\alpha$ and $\beta$, which are independent and equally likely. There are three social choice options, called $x$, $y$ and $z$. Each individual's utility for these options depends on his type according to the following table.

| *Option* | $1, \alpha$ | $1, \beta$ | $2, \alpha$ | $2, \beta$ |
|---|---|---|---|---|
| $x$ | 2 | 0 | 2 | 2 |
| $y$ | 1 | 4 | 1 | 1 |
| $z$ | 0 | 9 | 0 | −8 |

The incentive-efficient mechanism that maximizes the *ex ante* expected sum of the two individuals' utilities is as follows: if 1 reports type $\alpha$ and 2 reports $\alpha$ then choose $x$, if 1 reports type $\beta$ and 2 reports $\alpha$ then choose $z$, and if 2 reports $\beta$ then choose $y$ (regardless of 1's report). However, Holmstrom and Myerson argue that such a mechanism would not be chosen in a mechanism-selection game that is played when 1 already knows his type, because, when 1 knows that his type is $\alpha$, he could do better by proposing to select the mechanism that always chooses $x$, and 2 would always want to accept this proposal. That is, because 1 would have no incentive to conceal his type from 2 in a mechanism-selection game if his type were $\alpha$ (when his interests would then have no conflict with 2's), we should not expect the individuals in a mechanism-selection game to agree inscrutably to an incentive-efficient mechanism that implicitly puts as much weight on 1's type-$\beta$ payoff as the mechanism described above.

For another example, consider again the sender–receiver game due to Farrell. Recall that $y$ would be the only possible equilibrium outcome if the individuals could communicate only face-to-face, with no mediation or other noise in their communication channel. Suppose that the mechanism-selection process is as follows: first 2 proposes a mediator who is committed to implement some incentive-compatible mechanism; then 1 can either accept this mediator and communicate with 2 thereafter only through him, or 1 can reject this mediator and thereafter communicate with 2 only face-to-face. Suppose now that 2 proposes that they should use a mediator who will implement the incentive-compatible mediation plan that is best for 2 (recommending $x$ with probability 2/3 and $y$ with probability 1/3 if 1 reports $\alpha$, recommending $y$ with probability 2/3 and $z$ with probability 1/3 if 1 reports $\beta$). We

have seen that this mechanism is worse than $y$ for 1 if his type is $\beta$. Furthermore, this mechanism would be worse than $y$ for player 1 under the *ex ante* welfare criterion, when his expected payoffs for type $\alpha$ and type $\beta$ are averaged, each with weight 1/2. However, it is an equilibrium of this mechanism-selection game for player 1 always to accept this proposal, no matter what his type is. If 1 rejected 2's proposed mediator, then 2 might reasonably infer that 1's type was $\beta$, in which case 2's rational choice would be $z$ instead of $y$, and $z$ is the worse possible outcome for both of 1's types.

Now consider a different mechanism-selection process for this example, in which the informed player 1 can select any incentive-compatible mechanism himself, with only the restriction that 2 must know what mechanism has been selected by 1. For any incentive-compatible mechanism $\mu$, there is an equilibrium in which 1 chooses $\mu$ for sure, no matter what his type is, and they thereafter play the honest and obedient equilibrium of this mechanism. To support such an equilibrium, it suffices to suppose that, if any mechanism other than $\mu$ were selected, then 2 would infer that 1's type was $\beta$ and therefore choose $z$. Thus, concepts like sequential equilibrium from non-cooperative game theory cannot determine the outcome of this mechanism-selection game, beyond what we already know from the revelation principle; we cannot even say that 1's selected mechanism will be incentive-efficient. To get incentive efficiency as a result of mechanism-selection games, we need some further assumptions, like those of cooperative game theory.

An attempt to extend traditional solution concepts from cooperative game theory to the problem of bargaining over mechanisms has been proposed by Myerson (1983; 1984a; 1984b). In making such an extension, one must consider not only the traditional problem of how to define reasonable compromises between the conflicting interests of different individuals, but also the problem of how to define reasonable compromises between the conflicting interests of different types of the same individual. That is, to conceal his type in the mechanism-selection process, an individual should bargain for some inscrutable compromise between what he really wants and what he would have wanted if his type had been different; and we need some formal theory to predict what a reasonable inscrutable compromise might be. In the above sender–receiver game, where only type $\beta$ of player 1 should feel any incentive to conceal his type, we might expect an inscrutable compromise to be resolved in favor of type $\alpha$. That is, in the mechanism-selection game where 1 selects the mechanism, we might expect both types of 1 to select the incentive-compatible mechanism that is best for type $\alpha$. (In this mechanism, the mediator recommends $x$ with probability 0.8 and $y$ with probability 0.2 if 1 reports $\alpha$; and the mediator recommends $x$ with probability 0.4, $y$ with probability 0.4, and $z$ with probability 0.2 if 1 reports $\beta$.) This mechanism is the *neutral optimum* for player 1, in the sense of Myerson (1983).

ROGER B. MYERSON

*See also* **incentive compatibility; mechanism design: new developments; revelation principle.**

*Bibliography*


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

Aumann, R.J. 1987. Correlated equilibrium as an expression of Bayesian rationality. *Econometrica* 55, 1–18.

Coase, R. 1960. The problem of social cost. *Journal of Law and Economics* 3, 1–44.

Cramton, P., Gibbons, R. and Klemperer, P. 1987. Dissolving a partnership efficiently. *Econometrica* 55, 615–32.

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

Farrell, J. 1993. Meaning and credibility in cheap-talk games. *Games and Economic Behavior* 5, 514–31. Repr. in *Mathematical Models in Economics*, ed. M. Bacharach and M. Dempster. Oxford: Oxford University Press, 1997.

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

Harris, M. and Raviv, A. 1981. Allocation mechanisms and the design of auctions. *Econometrica* 49, 1477–99.

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

Harsanyi, J.C. 1967. Games with incomplete information played by Bayesian players. *Management Science* 14, 159–82, 320–34, 481–502.

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

Holmstrom, B. and Myerson, R.B. 1983. Efficient and durable decision rules with incomplete information. *Econometrica* 51, 1799–19.

Lewis, T.R. and Sappington, D.E.M. 1989. Countervailing incentives in agency problems. *Journal of Economic Theory* 49, 294–313.

Myerson, R.B. 1979. Incentive compatibility and the bargaining problem. *Econometrica* 47, 61–74.

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

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

Myerson, R.B. 1983. Mechanism design by an informed principal. *Econometrica* 51, 1767–97.

Myerson, R.B. 1984a. Two-person bargaining problems with incomplete information. *Econometrica* 52, 461–87.

Myerson, R.B. 1984b. Cooperative games with incomplete information. *International Journal of Game Theory* 13, 69–86.

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

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

Myerson, R.B. 1988. Incentive constraints and optimal communication systems. In *Proceedings of the Second Conference on Theoretical Aspects of Reasoning about Knowledge*, ed. M.Y. Vardi. Los Altos: Morgan Kaufmann.

Myerson, R.B. and Satterthwaite, M. 1983. Efficient mechanisms for bilateral trading. *Journal of Economic Theory* 29, 265–81.

Ortega-Reichert, A. 1968. Models for competitive bidding under uncertainty. Ph.D. thesis, Department of Operations Research, Stanford University.

Palfrey, T. and Srivastava, S. 1987. On Bayesian implementable allocations. *Review of Economic Studies* 54, 193–208.

Riley, J.G. and Samuelson, W.F. 1981. Optimal auctions. *American Economic Review* 71, 381–92.