Mechanism Design for the Environment

Sandeep Baliga
Kellogg Graduate School of Management (M.E.D.S.),
Northwestern University

Eric Maskin
Institute for Advanced Study and Princeton University

September 11, 2002

1 Introduction

Economists are accustomed to letting the "market" solve resource-allocation problems. The primary theoretical justification for this laissez-faire position is the "first fundamental theorem of welfare economics" (see Debreu (1957)), which establishes that, provided all goods are priced, a competitive equilibrium is Pareto efficient. Implicit in the "all-goods-priced" hypothesis, however, is the assumption that there are no significant externalities; an externality, after all, can be thought of as an unpriced commodity.

Once externalities are admitted, the first welfare theorem no longer applies. Thus, a school of thought dating back to Pigou (1932), if not earlier, calls for government-imposed "mechanisms" (e.g., taxes on pollution) as a way of redressing the market failure.[1]

In opposition to the Pigouvian school, however, proponents of the Coase Theorem (Coase, 1960) argue that, even in the presence of externalities, economic agents should still be able to ensure a Pareto-efficient outcome without government intervention provided that there are no constraints on their ability to bargain and contract. The argument is straightforward: if a prospective allocation is inefficient, agents will have the incentive to bargain their way to a Pareto improvement. Thus, even if markets themselves fail, Coasians hold that there is still a case for laissez-faire.

The Coasian position depends, however, on the requirement that any externality present be *excludable* in the sense that the agent giving rise to it

has control over who is and who is not affected by it. A pure public good, which, once created, will be enjoyed by everybody, constitutes the classic example of a nonexcludable externality.[2]

To see what goes wrong with nonexcludable externalities, consider pollution. For many sorts of pollution, particularly that of the atmosphere or sea, it is fairly accurate to say that a polluter cannot choose to pollute one group of agents rather then another, that is, pollution can be thought of as a pure public bad and hence pollution reduction as a public good.

Now imagine that there is a set of communities that all emit pollution and are adversely affected by these emissions. Suppose, however, that reducing pollution emission is costly to a community (say, because it entails curtailing or modifying the community's normal activities). It is clear that if communities act entirely on their own, there will be too little pollution reduction, since a community shares the benefit of its reduction with the other communities but must bear the full cost alone. A Coasian might hope, however, that if communities came together to negotiate a pollution-reduction agreement – in which each community agrees to undertake some reduction in exchange for other communities' promises to do the same – a Pareto-efficient reduction might be attainable. The problem is, however, that any given community (let us call it "C") will calculate that if all the other communities negotiate an agreement, it is better off not participating. By staying out, C can enjoy the full benefits of the negotiated reduction (this is where the nonexcludibility assumption is crucial) without incurring any of the cost. Presumably, the

agreed reduction will be somewhat smaller than had C participated (since the benefits are being shared among only $N-1$ rather then N participants). However, this difference is likely to be small relative to the considerable saving to C from not bearing any reduction costs (we formalize this argument in section 2 below).[3]

Hence, it will pay community C to *free-ride* on the others' agreement. But since this is true for *every* community, there will end up being no pollution-reduction agreement at all, i.e., the only reduction undertaken will be on an individual basis. We conclude that, in the case of nonexcludable public goods, even a diehard Coasian should agree that outside intervention is needed to achieve optimality. The government - or some other coercive authority - must be called on to *impose* a method for determining pollution reduction. We call such a method a *mechanism* (or game form). Devising a suitable mechanism may, however, be complicated by the fact that the authority might not know critical parameters of the problem (e.g., the potential benefits that different communities enjoy from pollution reduction).

Because environmental issues often entail nonexcludable externalities, the theory of mechanism design (sometimes called "implementation theory") is particularly pertinent to the economics of the environment. In this short survey, we review some of the major concepts, ideas, and findings of the mechanism-design literature and their relevance for the environment.

We necessarily focus on only a few topics from a vast field. Those interested in going further into the literature are referred to the following other

surveys and textbooks: Corchon (1996), Chapter 7 of Fudenberg and Tirole (1991), Groves and Ledyard (1987), Jackson (2001) and (2001a), Laffont and Martimort (2002), Maskin (1985), Maskin and Sjöström (2001), Moore (1992), Chapters 6 and 10 of Myerson (1991), Palfrey (1992) and (2001).

2 The Model

There are N players or *agents,* indexed by $j \in \{1, 2, .., N\}$, and a set of *social choices (*or *social decisions)* Y with generic element y. Agents have preferences over the social choices, and these depend on their preference parameters or *types*. Agent j of type $\theta_j \in \Theta_j$ has a utility function $U_j(y, \theta_j)$ (the interpretation of agent j as a firm is one possibility, in which case U_j is firm j's profit function). Let $\theta \equiv (\theta_1, .., \theta_N) \in \Theta \equiv \Pi_{i=1}^{N} \Theta_i$ be the *preference profile* or state. A choice y is *(ex-post) Pareto-efficient* for preference profile θ if there exists no other decision y' such that, for all $i = 1, ..., N$,

$$U_i(y', \theta_i) \geq U_i(y, \theta_i)$$

with strict inequality for some i. A *social choice function (or decision rule)* f is a rule that prescribes an appropriate social choice for each state, i.e., a mapping $f : \Theta \to Y$. We say that f is *efficient* if $f(\theta)$ is Pareto efficient in each state θ.

We illustrate this set-up with an example based on the discussion of pollution in the Introduction. Suppose that N communities (labelled $i = 1, ..., N$) would like to reduce their aggregate emission of pollution. Suppose

that the gross benefit to community j of a pollution reduction r is $\theta_j\sqrt{r}$ where $\theta_j \in [a, b]$, and that the cost per unit of reduction is 1. If r_j is the reduction of pollution by community j, $r = \sum_{i=1}^{N} r_i$, and t_j is a monetary transfer to community j, then an social choice y takes the form $y = (r_1, ..., r_N, t_1, ..., t_N)$, and

$$U_j(y, \theta_j) = \theta_j\sqrt{r} - r_j + t_j.$$

We will assume that there is no net source of funds for the N agents, and so for *feasibility* it must be the case that

$$\sum_{i=1}^{N} t_i \leq 0.$$

The stronger requirement of *balance* entails that

$$\sum_{i=1}^{N} t_i = 0.$$

To see why Coasian bargaining will *not* lead to Pareto-efficient pollution reduction, observe first that because preferences are quasi-linear, any efficient social choice function that does not entail infinite transfers (either positive or negative) to some communities must implicitly place equal weight on all communities. Hence, the Pareto-efficient reduction $r^*(\theta_1, ..., \theta_N)$ will maximize

$$(\sum_{i=1}^{N} \theta_i)\sqrt{r} - r,$$

and so

$$r^*(\theta_i, ..., \theta_n) = \frac{(\sum \theta_i)^2}{4}. \tag{1}$$

However, if there is no reduction agreement, community j will choose $r_j = r_j^{**}(\theta_j)$ to maximize $\theta_j \sqrt{r_j + \sum_{i \neq j} r_i(\theta_i)} - r_j$. Thus, if none of the θ_i's are equal, we have

$$r_j^{**}(\theta_j) = \begin{cases} \theta_j^2/4 & \text{, if } \theta_j \text{ is maximal in } \{\theta_1, ..., \theta_n\} \\ 0 & \text{, otherwise} \end{cases}$$

and so the total reduction is

$$r^{**}(\theta_1, ..., \theta_n) = \sum_{i=1}^{N} r_i^{**}(\theta_i) = \max_j \frac{\theta_j^2}{4}. \tag{2}$$

Note the sharp contrast between (1) and (2). In particular, if all the θ_i's are in a small neighborhood of z, then (1) reduces approximately to $\frac{n^2 z^2}{4}$, whereas (2) becomes $\frac{z^2}{4}$. In other words, the optimum reduction differs from the reduction that will actually occur by a factor $\frac{n^2}{4}$.

Now, suppose that the communities attempt to negotiate the Pareto-efficient reduction (1) by, say, agreeing to share the costs in proportion to their benefits. That is, community j will pay a cost equal to $\frac{\theta_j \sum_{i=1}^{N} \theta_i}{4}$, so that its net payoff is

$$\theta_j \sqrt{\frac{\left(\sum \theta_i\right)^2}{4}} - \frac{\theta_j \left(\sum_{i=1}^{N} \theta_i\right)}{4} = \frac{\theta_j \left(\sum_{i=1}^{N} \theta_i\right)}{4}. \tag{3}$$

If instead, however, community j stands back and lets the others undertake the negotiation and costs, it will enjoy a pollution reduction of

$$r^*(\theta_{-j}) = \frac{\left(\sum_{i \neq j} \theta_i\right)^2}{4}$$

and, hence, realize a net payoff of

$$\frac{\theta_j \left(\sum_{i \neq j} \theta_i \right)}{2}. \tag{4}$$

But provided that

$$\sum_{i \neq j} \theta_i > \theta_j, \tag{5}$$

(4) exceeds (3), and so community j does better to free-ride on the others' agreement. Furthermore, as we have assumed that all the θ_i's are distinct, notice that (5) must hold for *some* j, and so a Pareto-efficient agreement is not possible. Indeed, the same argument shows that any agreement involving two or more communities is vulnerable to free-riding. Thus, despite the possibility of negotiation, pollution reduction turns out to be no greater than in the case where negotiation is ruled out.

We conclude that some sort of government intervention is called for. Probably the simplest intervention is for the government to impose a vector of quotas $(q_1, ..., q_N)$, where for each j, community j is required to reduce pollution by at least the amount q_j. If $q_j = \dfrac{\theta_j \left(\sum_{i=1}^{N} \theta_i \right)}{4}$, then the resulting outcome will be Pareto efficient.

Another familiar kind of intervention is for the government to set a vector of subsidies $(s_1, ..., s_N)$, where, for each j, community j is paid s_j for each unit by which it reduces pollution (actually this is not quite complete: to finance the subsidies - and thereby ensure feasibility - each community must also be taxed some fixed amount). If $s_j = 1 - \dfrac{\theta_j}{\sum_{i=1}^{N} \theta_i}$, then the outcome

induced by the subsidies will be Pareto efficient.

Notice that both these solutions rely on the assumption that the state is verifiable to the government.[4]But the more interesting - and typically harder - case is the one in which the preference profile is not verifiable. In that case, there are two particular information environments that have been most intensely studied: first, the preference profile could, although unobservable to the government, be observable to all the agents (*complete information*); or, second, each agent j could observe only his *own* preference parameter θ_j (*incomplete information*). In either case, the government typically "elicits" the true state by having the agents play a game or mechanism.

Formally, a *mechanism* is a pair (M, g) where M_i is agent i's *message space*, $M = \Pi_{i=1}^{N} M_i$ is the product of the individual message spaces with generic element m, $g : M \rightarrow Y$ is an *outcome function*, and $g(m) \in Y$ is the social choice.

Returning to our pollution example, we note that if each community j observes only its own type θ_j, the government might have the community "announce" its type so that $M_j = \Theta_j$. As a function of the profile of their announcements $\hat{\theta}$,[5]the government chooses the reduction levels and transfers:

$$g(\hat{\theta}) = (r_1(\hat{\theta}), ..., r_N(\hat{\theta}), t_1(\hat{\theta}), ...t_N(\hat{\theta})).$$

To predict the outcome of the mechanism, we must invoke an equilibrium concept. Because which equilibrium concept is appropriate depends on the information environment, we study the complete and incomplete information

settings separately.

3 Complete Information

We begin with complete information. This is the case in which all agents observe the preference profile (the state) θ but it is unverifiable to the mechanism-imposing authority. It is most likely to be a good approximation when the agents all know one another well, but the authority is a comparative outsider.

Let S be a equilibrium concept such as Nash equilibrium, subgame perfect equilibrium, etc. Let $O_S(M, g, \theta)$ be the set of equilibrium outcomes of mechanism (M, g) in state θ.

A social choice function f is *implemented by the mechanism* (M, g) in the solution concept S if $O_S(M, g, \theta) = f(\theta)$ for all $\theta \in \Theta$. In that case, we say f is *implementable* in S. Notice that, in every state, we require that *all* the equilibrium outcomes be optimal (we will say more about this below).

3.1 Nash Implementation

Suppose first that S is Nash equilibrium. A message profile m is a *Nash equilibrium* in state θ if

$$U_i(y(m), \theta_i) \geq U_i(y(m_i', m_{-i}), \theta_i)$$

for all $i = 1, ..., N$, and all $m_i' \in M_i$ where m_{-i} is the profile of messages $(m_1, ..., m_{i-1}, m_{i+1}, ..., m_N)$ that excludes m_i.

We note that it is easy to ensure that at least one equilibrium outcome coincides with what the social choice function prescribes if there are three or more agents ($N \geq 3$): let all agents announce a state simultaneously. If $N-1$ or more agree and announce the same state $\hat{\theta}$, then let $g(\hat{\theta}) = f(\hat{\theta})$; define the outcome arbitrarily if fewer than $N-1$ agents agree. Notice that, if θ is the true state, it is an equilibrium for every agent to announce $\hat{\theta} = \theta$, leading to the outcome $f(\theta)$, since a unilateral deviation by any single agent will not change the outcome. However, it is equally well an equilibrium for agents to unanimously announce any other state (and there are many nonunanimous equilibria as well). Hence, uniqueness of the equilibrium outcome is a valuable property of an implementing mechanism.

To ensure that it is possible to construct such a mechanism, we require the social choice function to satisfy monotonicity. A social choice function f is *monotonic* if for any $\theta, \phi \in \Theta$ and $y = f(\theta)$ such that $y \neq f(\phi)$, there exists an agent i and outcome y' such that $U_i(y, \theta_i) \geq U_i(y', \theta_i)$ but $U_i(y', \phi_i) > U_i(y, \phi_i)$. That is, a social choice function is monotonic if whenever there is an outcome y that is optimal in one state θ but not in another ϕ, there exists an agent i and an outcome y' such that agent i strictly prefers y' to y in state ϕ but weakly prefers y to y' in state θ. This is a form of "preference reversal."

The other condition on social choice functions we will impose to guarantee implementability is no veto power. A social choice function f satisfies *no veto power* if whenever agent i, state θ and outcome y are such that $U_j(y, \theta_j) \geq$

$U_j(y', \theta_j)$ for all agents $j \neq i$ and all $y' \in y$, then $y = f(\theta)$. That is, if in state θ, $N-1$ or more agents agree that the best possible outcome is y, then y is prescribed by f in state θ. Notice that in our pollution example, there is *no* alternative that any agent thinks is best: an agent would always prefer a bigger monetary transfer. Hence, no veto power is automatically satisfied.

Theorem 1 *(Maskin (1999)) If a social choice function is implementable in Nash equilibrium, then it is monotonic. If $N \geq 3$, a social choice function that satisfies monotonicity and no veto power is Nash implementable.*

Proof. Necessity: Suppose f is Nash implementable using the mechanism (M, g). Suppose m is a Nash equilibrium of (M, g) in state θ, where $f(\theta) = y$. Then, $g(m) = y$. But, if $f(\theta) \neq f(\phi)$, m cannot be a Nash equilibrium in state ϕ. Therefore, there must exist an agent i with a message m_i' and an outcome $y' = g(m_i', m_{-i})$ such that

$$U_i(y', \phi_i) = U_i(g(m_i', m_{-i}), \phi_i) > U_i(g(m), \phi_i) = U_i(y, \phi_i).$$

But because m is a Nash equilibrium in state θ, agent i must be willing to send the message m_i rather than m_i' in state θ. Hence,

$$U_i(y, \theta_i) \geq U_i(y', \theta_i),$$

implying that f is monotonic.

Sufficiency: See Maskin (1999). ■

It is not hard to verify that in out pollution example, the efficient social choice function $f(\theta) = (r_1(\theta), ..., r_N(\theta), t_1(\theta), ...t_N(\theta))$, where, for all j,

$$r_j(\theta) = \frac{\theta_j \sum_{i=1}^{N} \theta_i}{4} \tag{6}$$

and

$$t_j(\theta) = 0, \tag{7}$$

is monotonic and hence Nash implementable. To see this, choose θ and θ', and let $y = (r_1, ..., r_N, t_1, ..., t_N) = f(\theta)$. Then, from (6) and (7), $r_j = \frac{\theta_j \sum_{i=1}^{N} \theta_i}{4}$ and $t_j = 0$ for all j. For concreteness, suppose that, for some j, $\theta_j < \theta'_j$. Note that

$$U_j(y, \theta_j) = \frac{\theta_j \sum_{i=1}^{N} \theta_i}{2} - \frac{\theta_j \sum_{i=1}^{N} \theta_i}{4}. \tag{8}$$

Choose $y' = (r'_1, ..., r'_N, t'_1, ..., t'_N)$ such that,

$$\sum_{i=1}^{N} r'_i = \left(\sum_{i=1}^{N} \theta_i \right)^2 \tag{9}$$

$$r'_j = r_j = \frac{\theta_j \sum_{i=1}^{N} \theta_i}{4} \tag{10}$$

and

$$t'_j = \frac{-\theta_j \sum_{i=1}^{N} \theta_i}{2}. \tag{11}$$

From (6)-(11), we have

$$U_j(y', \theta_j) = U_j(y, \theta_j).$$

But because $\theta'_j > \theta_j$ *and* $\sum_{i=1}^{N} r'_i > \sum_{i=1}^{N} r_i$ we have

$$U_j(y', \theta'_j) > U_j(y, \theta'_j),$$

as monotonicity requires.

Here is an alternative but equivalent definition of monotonicity: A social choice function is *monotonic* if, for any θ,ϕ, and $y = f(\theta)$ such that

$$U_i(y, \theta_i) \geq U_i(y, \theta_i) \Rightarrow U_i(y, \phi_i) \geq U_i(y, \phi_i) \text{ for all } i,$$

we have $y = f(\phi)$. This rendition of monotonicity says that when the outcome that was optimal in state θ goes up in everyone's preference ordering when the state becomes ϕ, then it must remain socially optimal. Although this may seem like a reasonable property, monotonicity can be quite a restrictive condition:

Theorem 2 *(Muller and Satterthwaite (1977)). Suppose that Θ consists of all strict preference orderings on the social choice space Y. Then, any social choice function that is monotonic and has a range including at least three choices is dictatorial (i.e., there exists an agent i^* such that in all states agent i^*'s favorite outcome is chosen).*[6]

3.2 Other Notions of Implementation

One way to relax monotonicity is to invoke *refinements* of Nash equilibrium, which make it easier to knock out unwanted equilibria while retaining optimal ones. Let us, in particular, explore the concept of *subgame perfect equilibrium* and the use of *sequential mechanisms*, i.e., mechanisms in which agents send messages one at a time. We maintain the assumption that the preference profile is common knowledge among the agents but is unverifiable by an

outside party. Therefore, we consider mechanisms of *perfect information* and (this is the subgame perfection requirement) strategies that constitute a Nash equilibrium at *any* point in the game.

Rather than stating general theorems, we focus immediately on our pollution example. For simplicity, restrict attention to the case of two communities ($N = 2$). We shall argue that *any* social choice function in this setting is implementable in subgame perfect equilibrium using a sequential mechanism.

We note first that, for $i = 1, 2$ and any $\theta_i, \theta_i' \in (a, b)$ there exist $(r_1^o(\theta_i, \theta_i'),$ $r_2^o(\theta_i, \theta_i'),$ $t_i^o(\theta_i, \theta_i'))$ and $(r_1^{oo}(\theta_i, \theta_i'),$ $r_2^{oo}(\theta_i, \theta_i'),$ $t_i^{oo}(\theta_i, \theta_i'))$ such that

$$\begin{aligned} &\theta_i \sqrt{r_1^o(\theta_i, \theta_i') + r_2^o(\theta_i, \theta_i')} - r_i^o(\theta_i, \theta_i') + t_i^o(\theta_i, \theta_i') \\ &> \theta_i \sqrt{r_1^{oo}(\theta_i, \theta_i') + r_2^{oo}(\theta_i, \theta_i')} - r_i^{oo}(\theta_i, \theta_i') + t_i^{oo}(\theta_i, \theta_i') \end{aligned} \tag{12}$$

and

$$\begin{aligned} &\theta_i' \sqrt{r_1^{oo}(\theta_i, \theta_i') + r_2^{oo}(\theta_i, \theta_i')} - r_i^{oo}(\theta_i, \theta_i') + t_i^{oo}(\theta_i, \theta_i') \\ &> \theta_i' \sqrt{r_1^o(\theta_i, \theta_i') + r_2^o(\theta_i, \theta_i')} - r_i^o(\theta_i, \theta_i') + t_i^o(\theta_i, \theta_i') \end{aligned} \tag{13}$$

Formulas (12) and (13) constitute a *preference reversal condition*. The condition says that for any two types θ_i and θ_i' we can find choices (r_1^o, r_2^o, t_i^o) and $(r_1^{oo}, r_2^{oo}, t_i^{oo})$ such that the former is preferred to the latter under θ_i and the latter is preferred to the former under θ_i'.

In view of preference reversal, we can use the following mechanism to implement a given social choice function f:

Stage 1

Stage 1.1: Agent 1 announces a type $\hat{\theta}_1$.

Stage 1.2: Agent 2 can *agree*, in which case we go to Stage 2, or disagree by announcing some $\hat{\theta}_1' \neq \hat{\theta}_1$, in which case we go to Stage 1.3.

Stage 1.3: Agent 1 is fined some large amount $p^*$and then chooses between $(r_1^o(\hat{\theta}_1, \hat{\theta}_1'), r_2^o(\hat{\theta}_1, \hat{\theta}_1'), t_1^o(\hat{\theta}_1, \hat{\theta}_i'))$ and $(r_1^{oo}(\hat{\theta}_1, \hat{\theta}_1'), r_2^{oo}(\hat{\theta}_1, \hat{\theta}_1'), t_1^{oo}(\hat{\theta}_1, \hat{\theta}_1'))$. If he chooses the former, agent 2 is also fined p^*; if he chooses the latter, agent 2 receives p^*. The mechanism stops here.

Stage 2: This is the same as Stage 1.2 except the roles are reversed: agent 2 announces $\hat{\theta}_2$, and agent 1 can either agree or disagree. If he agrees, we go to Stage 3. If he disagrees, then agent 2 is fined $p^*$and must choose between $(r_1^o(\hat{\theta}_2, \hat{\theta}_2'), r_2^o(\hat{\theta}_2, \hat{\theta}_2'), t_2^o(\hat{\theta}_2, \hat{\theta}_2'))$ and $(r_1^{oo}(\hat{\theta}_2, \hat{\theta}_2'), r_2^{oo}(\hat{\theta}_2, \hat{\theta}_2'), t_2^{oo}(\hat{\theta}_2, \hat{\theta}_2'))$. If he chooses the former, agent 1 is also fined p^*; if he chooses the latter, agent 1 receives p^*.

Stage 3: If $\hat{\theta}_1$ and $\hat{\theta}_2$ have been announced, the outcome $f(\hat{\theta}_1, \hat{\theta}_2)$ is implemented.

We claim that, in state (θ_1, θ_2), there is a unique subgame perfect equilibrium of this mechanism, in which agent 1 truthfully announces $\hat{\theta}_1 = \theta_1$ and agent 2 truthfully announces $\hat{\theta}_2 = \theta_2$, so that the equilibrium outcome is $f(\hat{\theta}_1, \hat{\theta}_2)$. To see this, note that in Stage 2, agent 1 has the incentive to disagree with any untruthful announcement $\hat{\theta}_2 \neq \theta_2$ by setting $\hat{\theta}_2' = \theta_2$. This is because agent 1 forecasts that, by definition of $(r_1^o(\hat{\theta}_2, \theta_2), r_2^o(\hat{\theta}_2, \theta_2), t_2^o(\hat{\theta}_2, \theta_2))$ and $(r_1^{oo}(\hat{\theta}_2, \theta_2), r_2^{oo}(\hat{\theta}_2, \theta_2), t_2^{oo}(\hat{\theta}_2, \theta_2))$ and from (13), agent 2 will choose the latter, and so 1 will collect the large sum p^*. By contrast, agent 1

will *not* disagree if $\hat{\theta}_2$ is truthful - i.e., $\hat{\theta}_2 = \theta_2$ - because otherwise (regardless of what $\hat{\theta}_2'$ he announces) (12) implies that agent 2 will choose $(r_1^o(\theta_2, \hat{\theta}_2'), r_2^o(\theta_2, \hat{\theta}_2'), t_2^o(\theta_2, \hat{\theta}_2'))$, thereby requiring 1 to pay a large fine himself. But this in turn means that agent 2 will announce truthfully because by doing so he can avoid the large fine that would be entailed by 1's disagreeing. Similarly, agent 1 will be truthful in Stage 1, and agent 2 will disagree if and only if 1 is untruthful. Because both agents are truthful in equilibrium, the desired outcome $f(\theta_1, \theta_2)$ results in Stage 3.

Herein we have examined only one simple example of implementation in a refinement of Nash equilibrium. For more thorough treatments, see the surveys by Moore(1992), Palfrey(2001), or Maskin and Sjöström(2001).

4 Incomplete Information

We next turn to incomplete information. This is the case in which agent i observes only his own type θ_i.

4.1 Dominant Strategies

A mechanism (M, g) that has the property that each agent has a dominant strategy - a strategy that is optimal regardless of the other agents' behavior - is clearly attractive since it means that an agent can determine his optimal message without having to calculate those of other agents, a calculation may be particularly complex under incomplete information.

Formally, a strategy μ_i for agent i is mapping from his type space Θ_i to

his message space M_i. A strategy, $\mu_i : \Theta_i \to M_i$, is *dominant* for type θ_i if:

$$U_i(g(\mu_i(\theta_i), m_{-i}), \theta_i) \geq U_i(g(m_i', m_{-i}), \theta_i)$$

for all $m_i' \in M_i$, $m_{-i} \in M_{-i}$. A strategy profile $\mu = (\mu_1, ..., \mu_N)$ is a *dominant strategy equilibrium* if, for all i and θ_i, $\mu_i(\theta_i)$ is dominant for θ_i.

A social choice function f is *implemented in dominant strategy equilibrium* by the mechanism (M, g) if there exists a dominant strategy equilibrium μ for which $g(\mu(\theta)) = f(\theta)$ for all $\theta \in \Theta$.[7]

Of course, implementation in dominant strategy equilibrium is a demanding requirement, and so perhaps not surprisingly it is difficult to attain in general:

Theorem 3 *(Gibbard(1973) and Satterthwaite(1975)) Suppose that Θ consists of all strict preference orderings. Then, any social choice function that is implementable in dominant-strategy equilibrium and whose range includes at least three choices is dictatorial.*

Proof. Suppose that f is implementable in dominant-strategy equilibrium and that the hypotheses of the theorem hold. Consider θ, $\theta' \in \Theta$ such that $f(\theta) = y$ and, for all i,

$$U_i(y, \theta_i) \geq U_i(y', \theta_i) \text{ implies } U_i(y, \theta_i') \geq U_i(y', \theta_i') \tag{14}$$

for all y'. By assumption, there exists a mechanism (M, g) with a dominant-strategy equilibrium μ such that $g(\mu(\theta)) = y$. We claim that

$$g(\mu(\theta')) = y. \tag{15}$$

To see why (15) holds, suppose that

$$g(\mu_1(\theta_1'), \mu_2(\theta_2), ..., \mu_N(\theta_N)) \neq g(\mu(\theta)) = y.$$

Then

$$U_1(g(\mu_1(\theta_1'), \mu_2(\theta_2), ..., \mu_N(\theta_N)), \theta_1')) > U_1(y, \theta_1'), \tag{16}$$

a contradiction of the assumption that $\mu_1(\theta_1)$ is dominant for θ_1. Hence,

$$g(\mu_1(\theta_1'), \mu_2(\theta_2), ..., \mu_N(\theta_N)) = y$$

after all. Continuing iteratively, we obtain

$$g(\mu_1(\theta_1'), \mu_2(\theta_2'), \mu_3(\theta_3), ..., \mu_N(\theta_N)) = y,$$

and

$$g(\mu(\theta')) = y. \tag{17}$$

But (17) implies that $f(\theta') = y$. We conclude that f is monotonic, and so Theorem 2 implies that it is dictatorial. ■

In contrast to the pessimism of Theorem 3, Vickrey (1961) and, more generally, Clarke (1971) and Groves (1973) have shown that much more positive results are obtainable when agents' preferences are quasi-linear. Specifically, suppose that we wish to implement a social choice function $f(\theta) = (r_1(\theta), ..., r_N(\theta), t_1(\theta), ..., t_N(\theta))$ entailing Pareto-efficient pollution reduction, i.e., such that

$$\sum_{i=1}^{N} r_i(\theta) = r^*(\theta), \tag{18}$$

where $r^*(\theta)$ solves

$$r^*(\theta) = \arg\max \sum_{i=1}^{N} \theta_i \sqrt{r} - r. \tag{19}$$

If community j is not allocated any transfer by the mechanism, then j solves

$$\max \theta_j \sqrt{\sum_{i \neq j} r_i + r_j} - r, \tag{20}$$

which clearly does not result in the total reduction being $r^*(\theta)$. To bring the maximands of individual communities and overall society into line, we shall give community j a transfer equal to the sum of the other communities' payoffs (net of transfers):

$$t_j(\hat{\theta}) = \sum_{i \neq j} (\hat{\theta}_i \sqrt{r^*(\hat{\theta})} - r_i(\hat{\theta})) + \tau_j(\hat{\theta}_{-j}), \tag{21}$$

where $\tau_j(.)$ is an arbitrary function of θ_{-j}. A mechanism in which each agent j announces $\hat{\theta}_j$ and the outcome is $(r_1(\hat{\theta}), ..., r_N(\hat{\theta}), t_1(\hat{\theta}), ..., t_N(\hat{\theta}))$ where $(r_1(.), ..., r_N(.))$ satisfies (18) and (19), and $(t_1(.), ..., t_N(.))$ satisfies (21), is called a *Groves scheme* (see Groves (1973)).

We claim that, in a Groves scheme, community j's telling the truth (announcing $\hat{\theta}_j = \theta_j$) is dominant for θ_j for all j and all θ_j. Observe that in such a mechanism, community j's overall payoff if it tells the truth and the other communities announce $\hat{\theta}_{-j}$ is

$$\begin{aligned} &\theta_j \sqrt{r^*(\theta_j, \hat{\theta}_{-j})} - r(\theta_j, \hat{\theta}_{-j}) + \sum_{i \neq j} (\hat{\theta}_i \sqrt{r^*(\theta_j, \hat{\theta}_{-j})} - r_i(\theta_j, \hat{\theta}_{-j})) + \tau_j(\hat{\theta}_{-j}) \\ &= (\theta_j + \sum_{i \neq j} \hat{\theta}_i) \sqrt{r^*(\theta_j, \hat{\theta}_{-j})} - r^*(\theta_j, \hat{\theta}_{-j}) + \tau_j(\hat{\theta}_{-j}). \end{aligned}$$

But from (19),

$$(\theta_j + \sum_{i \neq j} \hat{\theta}_i)\sqrt{r^*(\theta_j, \hat{\theta}_{-j})} - r^*(\theta_j, \hat{\theta}_{-j}) + \tau_j(\hat{\theta}_{-j}) \tag{22}$$
$$\geq (\theta_j + \sum_{i \neq j} \hat{\theta}_i)\sqrt{r'} - r' + \tau_j(\hat{\theta}_{-j})$$

for all r'. In particular, (22) holds when $r' = r^*(\hat{\theta}_j, \hat{\theta}_{-j})$, which then implies that taking $\hat{\theta}_j = \theta_j$ is dominant as claimed.

Thus, with one proviso, a Groves scheme succeeds in implementing the Pareto-efficient pollution reduction. The proviso is that we have not yet ensured that the transfer functions (21) are feasible. One way of ensuring feasibility is to take

$$\tau_j(\hat{\theta}_{-j}) = -\max_r \sum_{i \neq j} (\theta_i \sqrt{r} - r)$$

for all j.

Then, community j's transfer becomes

$$t_j(\hat{\theta}) = \sum_{i \neq j} (\hat{\theta}_i \sqrt{r^*(\hat{\theta})} - r_i(\hat{\theta}_i)) - \max_r (\sum_{i \neq j} \widehat{\theta}_i \sqrt{r} - r). \tag{23}$$

When transfers take the form (23), a Groves scheme is called a *pivotal mechanism* or a *Vickrey-Clarke-Groves mechanism*. Notice that the transfer (23) is always (weakly) negative, ensuring feasibility.

The logic underlying (23) is straightforward. If community j's announcement has no effect on the social choice, the community pays nothing. However, if it *does* change this choice (i.e., it is "pivotal"), j pays the corresponding loss imposed on the rest of society. Although the pivotal mechanism

is feasible, it is not balanced, i.e., the transfers do not sum to zero. Indeed, as shown by Green and Laffont (1979), *no* Groves scheme is balanced. Furthermore, arguments due to Green and Laffont (1977) imply that in a slightly more general version of our pollution example, Groves schemes are essentially the *only* mechanisms that implement social choice functions with Pareto-efficient pollution reductions. This motivates the search for balanced mechanisms that invoke a less demanding notion of implementation than in dominant-strategy equilibrium, a question we turn to in the next subsection.

We have been assuming that each community j's payoff depends directly only on its own preference parameter θ_j. Radner and Williams (1988) extend the analysis to the case when j's payoff may depend on the entire profile θ. We have also been concentrating on the case of *Pareto-efficient* social choice functions (or at least social choice functions for which the pollution reduction is Pareto-efficient); Dasgupta, Hammond, and Maskin (1980) examine dominant-strategy implementation of more general social choice functions.

4.2 Bayesian Equilibrium

Dominant-strategy equilibrium requires that each agent be willing to use his equilibrium strategy whatever the behavior of the other agents. Bayesian equilibrium requires only that each agent be willing to use his equilibrium strategy when he expects other agents to do the same. A couple of points are worth noting here. First, because agents' equilibrium strategies depend on their types but, given the incomplete information, an agent does not

know others' types, we must specify his *beliefs* about these types to complete the description of the model. Second, if a social choice function is implementable in dominant-strategy equilibrium, then it is certainly implementable in Bayesian equilibrium, so by moving to the latter concept, we are weakening the notion of implementation.

We assume that agents' types are independently distributed; the density and distribution functions for agent i of type $\theta_i \in [a, b]$ are $p_i(\theta_i)$ and $P_i(\theta_i)$ respectively. We suppose that these distributions are common knowledge amongst the agents. Hence, the c.d.f. for agent i's beliefs over the types of the other agents is given by $F_i(\theta_{-i}) \equiv \prod_{j \neq i} P_j(\theta_j)$.

There are two critical conditions that a social choice function must satisfy to ensure that it is implementable in Bayesian equilibrium (see Postlewaite and Schmeidler (1986), Palfrey and Srivastava (1989) and Jackson (1991)). The first is Bayesian incentive-compatibility. A social choice function f is *Bayesian incentive compatible (BIC)* if

$$E_{\theta_{-i}}[U_i(f(\theta_i, \theta_{-i}), \theta_i)] \geq E_{\theta_{-i}}[U_i(f(\theta_i', \theta_{-i}), \theta_i)]$$

for all i, and $\theta_i, \theta_i' \in \Theta_i$, where

$$E_{\theta_{-i}}[U_i(f(\theta_i, \theta_{-i}), \theta_i)] = \int_{\Theta_{-i}} U_i(f(\theta_i, \theta_{-i}), \theta_i) dF_i(\theta_{-i}).$$

The second condition is the incomplete-information counterpart to monotonicity. For this purpose, we define a *deception for agent* j to be a function $\alpha_j : \Theta_j \to \Theta_j$. A *deception* α is a profile $\alpha = (\alpha_1, ..., \alpha_N)$. A social choice

function f is *Bayesian monotonic* if for all deceptions α such that $f \circ \alpha \neq f$ there exist j and a function $\gamma : \Theta_{-j} \to Y$ such that

$$EU_j(f(\theta_j, \theta_{-j}), \theta_j) \geq EU_j(\gamma(\theta_{-j}), \theta_j)$$

for all $\theta_j \in \Theta_j$, and

$$EU_j(f(\alpha(\theta'_j, \theta_{-j}), \theta'_j) < EU_j(\gamma(\alpha_{-j}(\theta_{-j})), \theta'_j)$$

for some $\theta'_j \in \Theta_j$.

Jackson (1991) shows that in quasi-linear settings, such as our pollution example, BIC and Bayesian monotonicity are not only necessary but sufficient for a social choice function to be implementable in Bayesian equilibrium.

Let us return to our pollution example. We noted in the previous subsection that a social choice function entailing Pareto-efficient pollution reduction (i.e., reduction satisfying (18) and (19)) cannot be implemented in dominant-strategy equilibrium if it is balanced. However, this negative conclusion no longer holds with Bayesian implementation.

To see this, consider a pollution reduction profile $(r_1^o(\theta), ..., r_N^o(\theta))$ that is Pareto-efficient (i.e., $\sum_{i=1}^N r_i^o(\theta) = r^*(\theta)$, where $r^*(.)$ satisfies (19)). Consider the mechanism in which each agent j announces $\hat{\theta}_j$ and the outcome is $(r_1^o(\hat{\theta}), ..., r_N^o(\hat{\theta}), t_1^o(\hat{\theta}), ..., t_N^o(\hat{\theta}))$, where $t_j^o(\hat{\theta})$ satisfies

$$t_j^o(\hat{\theta}) = \int_{\Theta_{-j}} \sum_{i \neq j} (\hat{x}_i \sqrt{r^*(\hat{\theta}_j, x_{-j})} - r_i(\hat{\theta}_j, x_{-j})) dF_j(x_{-j}) \tag{24}$$
$$- \frac{1}{N-1} \sum_{i \neq j} \int_{\Theta_{-i}} \sum_{k \neq i} (x_k \sqrt{r^*(\hat{\theta}_i, x_{-i})} - r_k(\hat{\theta}_i, x_{-i})) dF_i(x_{-i}).$$

Notice that the first term (integral) on the right-hand side of (24) is just the expectation of the sum in (21). Furthermore the other terms in (24) do not depend on $\hat{\theta}_j$. Hence, this mechanism can be thought of as an "expected Groves scheme." It was first proposed by Arrow (1979) and d'Aspremont and Gérard-Varet (1979).

The terms after the first integral in (24) are present to ensure balance. If all communities tell the truth (we verify below that the social choice function satisfies BIC $f(\theta) = (r_1^o(\hat{\theta}), ..., r_N^o(\hat{\theta}), t_1^o(\hat{\theta}), ..., t_N^o(\hat{\theta})))$, then observe that

$$\begin{aligned}
\sum_{j=1}^N t_j^o(\theta) &= \sum_{j=1}^N \int_{\Theta_{-j}} \sum_{i \neq j} (\theta_i \sqrt{r^*(\theta_j, \theta_{-j})} - r_i^o(\theta_j, \theta_{-j})) dF_j(\theta_{-j}) \\
&\quad - \frac{1}{N-1} \sum_{j=1}^N \sum_{i \neq j} \int_{\Theta_{-i}} \sum_{k \neq i} (\theta_k \sqrt{r^*(\theta_i, \theta_{-i})} - r_k^o(\theta_i, \theta_{-i})) dF_i(\theta_{-i}) \\
&= \sum_{j=1}^N \int_{\Theta_{-j}} \sum_{i \neq j} (\theta_i \sqrt{r^*(\theta_j, \theta_{-j})} - r_i^o(\theta_j, \theta_{-j})) dF_j(\theta_{-j}) \\
&\quad - \sum_{j=1}^N \int_{\Theta_{-i}} \sum_{i \neq j} (\theta_i \sqrt{r^*(\theta_j, \theta_{-j})} - r_i^o(\theta_j, \theta_{-j})) dF_j(\theta_{-j}) \\
&= 0,
\end{aligned}$$

as desired.

To see that BIC holds (so that truth-telling is an equilibrium) note that if $f(\theta) = (r_1^o(\theta), ..., r_N^o(\theta), t_1^o(\theta), ..., t_N^o(\theta))$, then, for all j, θ_j, θ'_j, and θ_{-j},

$$\begin{aligned} & E_{\theta_{-j}}[U_j(f(\theta'_j, \theta_{-j}), \theta_j)] \qquad (25) \\ & = E_{\theta_{-j}}[\theta_j \sqrt{r^*(\theta'_j, \theta_{-j})} - r_j^o(\theta'_j, \theta_{-j}) + t_j^o(\theta'_j, \theta_{-j})] \\ & = E_{\theta_{-j}}[\theta_j \sqrt{r^*(\theta'_j, \theta_{-j})} - r_j^o(\theta'_j, \theta_{-j}) \\ & \qquad + E_{\theta_{-j}} \sum_{i \neq j} (\theta_i \sqrt{r^*(\theta'_j, \theta_{-j})} - r_i^o(\theta'_j, \theta_{-j}))], \end{aligned}$$

where the last line of the right-hand side of (25) corresponds to the first term of $t_j^o(\theta'_j, \theta_{-j})$ as given by the right-hand side of (24), but with all but the first term omitted (since the other terms on the right-hand side of (24) do not depend on θ'_j and hence do not affect incentive compatibility for community j). But the last line of the right-hand side of (25) can be rewritten as

$$E_{\theta_{-j}}[(\sum_{j=1}^{N} \theta_i) \sqrt{r^*(\theta'_j, \theta_{-j})} - r^*(\theta'_j, \theta_{-j})]. \qquad (26)$$

By definition of $r^*(\theta)$, the square-bracketed expression in (26) is maximized when $\theta'_j = \theta_j$. Hence from (25) and (26), we have

$$E_{\theta_{-j}}[U_j(f(\theta'_j, \theta_{-j}), \theta_j)] \leq E_{\theta_{-j}}[U_j(f(\theta_j, \theta_{-j}), \theta_j)],$$

as required for BIC.

One can readily show that f also satisfies Bayesian monotonicity (but we will refrain from doing so here). Hence, we conclude that it is implemented by the Groves mechanism (actually, it turns out that the equilibrium outcome of the expected Groves mechanism is not unique, so, without modification,

that mechanism does not actually implement f). Thus relaxing the notion of implementability from dominant-strategy to Bayesian equilibrium permits the implementation of balanced social choice functions. On the downside, however, note that the very construction of the expected Groves mechanism requires common knowledge of the distribution of θ.

Footnotes

1. For more on Pigouvian taxes and other regulatory responses to pollution externalities, see the chapter in this Handbook by Gloria Helfand, Peter Berck, and Tim Maull, titled "The theory of pollution policy."

2. For more on the theory of externalities and public goods, see the chapter in this Handbook by David Starrett, titled "Property rights, public goods, and the environment."

3. Implicit in this argument is the assumption that the other communities cannot, in effect, *coerce* community C's participation by threatening, say, to refrain from negotiating any agreement at all if C fails to participate. What we have in mind is the idea that any such threat would not be credible, i.e., it would not actually be carried out if push came to shove. Also implicit is the presumption that community C will not be offered especially favorable terms in order to persuade it to join. But notice that if communities anticipated getting especially attractive offers by staying out of agreements, then they would *all* have the incentive to drag their heels about negotiating such agreements and so the same conclusion about the inadequacy of relying on

negotiated settlements would obtain. For further discussion of these points see Maskin (1994) and Baliga and Maskin(2002).

4. They also depend on the assumption that each community's reduction is verifiable. If only a noisy signal of a reduction is verifiable, then there is said to be moral hazard. However, we will assume throughout that the social choice is indeed verifiable so that the issue of moral hazard does not arise.

5. We write the profile of *announced* parameters as $\hat{\theta}$, to distinguish it from the *actual* parameters θ.

6. Monotonicity is a good deal less restrictive if one considers implementation of social choice *correspondences* rather than functions (see Maskin (1999)).

7. Notice that, unlike with implementation in Nash equilibrium, we require only that *some* dominant strategy equilibrium outcome coincide with $f(\theta)$, rather then that there be a unique equilibrium outcome. However, multiple equilibria are not typically a serious problem with dominant strategies. In particular, when preferences are *strict* (i.e., indifference is ruled out), the dominant-strategy equilibrium outcome is, indeed, unique.

Acknowledgments

We would like to thank Jeffrey Vincent, Karl-Goran Maler, David Starrett and Theodore Groves for their comments on an earlier version of this chapter.

References

Arrow, K. (1979), "The Property Rights Doctrine and Demand Revelation under Incomplete Information," in: M. Boskin, ed., *Economies and Human Welfare* (Academic Press, New York).

Baliga, S. and E. Maskin (2002), "The Free-Rider Problem and the Coase Theorem," work in progress.

Clarke, E. (1971), "Multi-Part Pricing of Public Goods," *Public Choice*, 11: 17-33.

Coase, R. (1960), "The Problem of Social Cost," *Journal of Law and Economics*, 3:1-44.

Corchon, L. (1996), *The Theory of Implementation of Socially Optimal Decisions in Economics*, New York: St. Martin's Press.

Dasgupta, P., P. Hammond, and E. Maskin (1980), "On Imperfect Information and Optimal Pollution Control," *Review of Economic Studies*, 47: 857-860.

d'Aspremont, C. and L.A. Gérard-Varet (1979), "Incentives and Incomplete Information, *Journal of Public Economics,* 11: 25-45.

Debreu, G. (1957), *Theory of Value*, New York: John Wiley & Sons.

Fudenberg, D. and J. Tirole (1991), *Game Theory* (MIT Press: Cambridge).

Green, J. and J.-J. Laffont (1977), "Characterization of Satisfactory Mechanism for the Revelation of Preferences for Public Goods," *Econometrica* 45: 727-738.

Green, J. and J.-J. Laffont (1979), *Incentives in Public Decision Making* (North Holland, Amsterdam).

Groves, T. (1973), "Incentives in Teams," *Econometrica,* 41: 617-663.

Groves, T. and J. Ledyard (1987), "Incentive Compatibility since 1972," in T. Groves, R. Radner, and S. Reiter, eds., *Information, Incentives and Economic Mechanisms*, Minneapolis: University of Minnesota Press.

Jackson, M. (1991), "Bayesian Implementation," *Econometrica*, 59: 461-477

Jackson, M. (2001), "Mechanism Theory," forthcoming in: *Encyclopedia of Life Support Systems.*

Jackson, M. (2001a), "A Crash Course in Implementation Theory," *Social Choice and Welfare*, 18: 655-708

Laffont, J.-J. and D. Martimort (2002), *The Theory of Incentives: The Principal-Agent Model*, Princeton: Princeton University Press.

Laffont, J.-J. and E. Maskin (1979), "A Differentiable Approach to Expected Utility Maximizing Mechanisms," in: J.-J. Laffont, ed., *Aggregation and Revelation of Preferences* (North-Holland, Amsterdam.), 289-308.

Mas-Colell, A., M. Whinston, and J. Green (1995), *Microeconomic Theory* (Oxford University Press, Oxford).

Maskin, E. (1985), "The Theory of Implementation in Nash Equilibrium: a Survey," in L. Hurwicz, D. Schmeidler, and H. Sonnenschein, eds., *Social Goals and Social Organization*, Cambridge: Cambridge University Press.

Maskin, E. (1994), "The Invisible Hand and Externalities," *American Economic Review*, 84(2): 333-337.

Maskin, E. (1999), "Nash Equilibrium and Welfare Optimality," *Review of Economic Studies,* 66: 23-38.

Maskin, E. and T. Sjöström (2001), "Implementation Theory," forthcoming in K.Arrow, A. Sen, and K. Suzumura, eds., *Handbook of Social Choice and Welfare*, Amsterdam: North Holland.

Moore, J. (1992), "Implementation in Environments with Complete Information," in: J.-J. Laffont , ed., *Advances in Economic Theory: Proceedings of the Sixth World Congress of the Econometric Society* (Cambridge University Press, Cambridge), 182-282.

Myerson, R. (1991), *Game Theory* (Harvard University Press, Cambridge).

Myerson, R. and M. Satterthwaite (1983), "Efficient Mechanisms for Bilateral Trade," *Journal of Economic Theory* 29: 265-281.

Palfrey, T. (1992), "Implementation in Bayesian Equilibrium: the Multiple Equilibrium Problem in Mechanism Design," in J.J. Laffont, ed., *Advances in Economic Theory*, Cambridge: Cambridge University Press

Palfrey, T. (2001), "Implementation Theory," forthcoming in R. Aumann and S. Hart, eds., *Handbook of Game Theory*, vol. 3, Mastermind: North-Holland

Palfrey, T. and S. Srivastava (1987), "On Bayesian Implementation Allocations," *Review of Economic Studies*, 54: 193-208.

Pigou, A.C. (1932), *The Economics of Welfare*, London: Macmillan.

Postlewaite, A. and D. Schmeidler (1986), "Implementation in Differential Information Economics," *Journal of Economic Theory*, 39: 14-33.

Radner, R. and S. Williams (1988), "Informational Externalities and the Scope of Efficient Dominant Strategy Mechanisms," mimeo.

Vickrey, W. (1961), "Counterspeculation, Auctions, and Competitive Sealed-Tenders," *Journal of Finance,* 16: 8-37.