The Theory of Implementation of Social Choice Rules[*]

Roberto Serrano[†]

Department of Economics, Box B
Brown University
Providence, RI 02912, U.S.A.

Working Paper No. 2003-19
September 2003

Abstract. Suppose that the goals of a society can be summarized in a social choice rule, i.e., a mapping from relevant underlying parameters to final outcomes. Typically, the underlying parameters (e.g., individual preferences) are private information to the agents in society. The implementation problem is then formulated: under what circumstances can one design a mechanism so that the private information is truthfully elicited and the social optimum ends up being implemented? In designing such a mechanism, appropriate incentives will have to be given to the agents so that they do not wish to misrepresent their information. The theory of implementation or mechanism design formalizes this "social engineering" problem and provides answers to the question just posed. I survey the theory of implementation in this article, emphasizing the results based on two behavioral assumptions for the agents (dominant strategies and Nash equilibrium). Examples discussed include voting, and the allocation of private and public goods under complete and incomplete information.

Journal of Economic Literature Classification: C72, D78, D82.

Keywords: Implementation theory, mechanism design, asymmetric information, decentralization, game theory, dominance, Nash equilibrium, monotonicity.

[†]Brown University, `roberto_serrano@brown.edu`; `www.econ.brown.edu/faculty/serrano`

1   Introduction

Suppose the goals of a group of agents, which we will refer to as a society, can be summarized in a mapping that prescribes the socially desirable outcomes as a function of certain underlying parameters. A first example consists of the allocation of goods via on-line auctions. The auctioneer may want to allocate the goods efficiently (i.e., assign them to those consumers that value them the most) or perhaps in a way that maximizes her expected revenue. Who should be allocated the goods and at what price will depend on the consumers' valuations for the goods, which are their private information. As a second example, suppose that a public project, such as a highway, is being proposed. When the public authority decides whether the project should be undertaken, she should be comparing its cost to the social surplus being created as a result of its construction (because of lower commuting time and an overall improvement in citizens' welfare), and perhaps fairness considerations should also be taken into account in the stipulation of the monetary payments that each agent will be asked to contribute (those who live near the new highway will be asked to pay less, for instance). Again, some essential aspects in this decision involve information privately held by the agents: apart from their individual valuations, the true cost of construction may be uncertain. A third class of examples is that of voting problems. A set of voters must elect a candidate for a political position. The set of social goals can be identified here with different political agendas (going from the left to the right of the political spectrum). Of course, each voter's individual political preference is also his private information. Finally, consider the renovation of a house. This will typically involve a contractual relationship between a home owner and a contractor. Its goal could be the efficient execution of the work, in terms of timing and quality of materials utilized. Several aspects of this relationship (e.g., how many days the plumber was expected to come but didn't show up, or how many phone calls the contractor made to find an appropriate plumber) will be impossible to verify by an outsider, such as a court of justice.[1]

Any of these examples can be characterized as an implementation problem. That is, the question is whether one can design a mechanism or institution where the agents will interact by taking actions in such a way that the socially desirable outcome is implemented, no matter what is the agents' private information. Let us appreciate why the problem is non-trivial. First,

---

[1]Some of these and other examples are presented in Subsection 2.1, after the necessary notation is introduced.

note that the private information will in general affect the social goal, i.e., the social choice is not a constant function. Second, agents' private preferences will have an impact on the way each of them behaves in the mechanism, i.e., what is best for each of them is also varying with their information. And third, as soon as they realize that they are interacting in the context of the mechanism, and prior to taking their actions, they will be making conjectures about what actions the other agents are likely to take, because every agent's action in principle will have an effect on the final outcome. These considerations make it clear that the methodology to be employed must involve game theory, which studies the behavior of rational strategic agents in situations of conflict.[2]

In formalizing the problem, we shall introduce the notion of an environment. An environment consists of a set of agents, a set of social alternatives or outcomes, agents' preferences over these outcomes, and the information possessed by each agent regarding the environment itself. To be clear, the mechanism designer (also called planner, outside enforcer, court of justice, etc.) is always at an informational disadvantage with respect to the agents, who, as a collective entity, know more about the true environment than does the designer. But in addition, one can consider complete information environments, in which every agent knows every relevant aspect of the environment (e.g., each other's preferences), or incomplete information environments, in which even agents may be asymmetrically informed. Aside from being the first step of analysis, complete information environments are also important in certain applications, typically involving a small number of agents: think of the contractual relationship described above, in which the exact events that have happened in the course of the renovations are commonly known by home owner and contractor.

The theory of implementation or mechanism design is concerned with the study and formalization of these "Social engineering" problems, and it provides answers to the important question of whether and how it is possible to implement different social choice rules. In this article, I will attempt to survey some of the central findings of implementation theory.[3] In doing so, I will confine my attention to static mechanisms, i.e., those in which each agent takes only one action, and at the time he takes it, he is unaware of the

---

[2]See Myerson (1991) and Osborne and Rubinstein (1994) for two excellent comprehensive books on the subject.

[3]For other surveys, the reader is referred to Maskin (1985), Moore (1992), Palfrey (1992, 2002), Corchón (1996), Jackson (2001) and Maskin and Sjostrom (2002). These are all excellent surveys for specialists. My attempt here is to reach out of the usual audience of implementation theory.

actions taken by the others. The theory combines its normative foundation, stemming from the relevance of the socially desirable goal, with a realistic positive component, because in designing the mechanism that will do the job, the "right" incentives will have to be given to the agents so that the desired outcome results from self-interested strategic behavior.

It is also important to observe that the theory does not rely on the existence of a mechanism designer who is someone other than the agents. Indeed, there my be no such entity, as in the case of the contract relationship. In cases like this, a contract (mechanism) can be designed by the two signing parties with an eye to providing the right incentives to both in pursuit of a commonly agreed goal. The contract will have to be written on the basis of verifiable information that, in the event of a dispute, an outside court will be able to settle. The role of the outside party here is not to design the mechanism, but simply to enforce it.

This survey will concentrate on two different modes of strategic behavior for the agents in the mechanism. First, we shall explore the possibilities of implementation in mechanisms where agents use dominant strategies. A strategy is dominant for an agent if it is always his best response, regardless of the actions taken by others. Whenever dominant strategies can be found in mechanisms, the corresponding implementation is quite robust because it is clear what each agent will end up doing, i.e., using his dominant strategy. The most important result of implementation in dominant strategies is the Gibbard-Satterthwaite theorem, which offers a striking negative result: essentially, if there are at least three social alternatives and the designer has very poor information so that she cannot rule out any of the possible agents' preferences in her design problem, the only social choice functions that can be implementable in dominant strategies are dictatorial: for these, the program of decentralizing decisions as done in a mechanism is defeated, because they could be imposed by a dictator.

To confront this bad news, the literature has proceeded in two different ways. First, it turns out that, if the designer is not completely in the dark so that she can impose certain restrictions on the preference domain, possibility results arise. In this case, it is crucial to identify the specific rules that can be implemented, because as was just pointed out, whenever one can implement in dominant strategies, one should do it. The second way out is based on considering a different behavioral assumption for the agents.

Indeed, one can argue that requiring dominance is too demanding. Perhaps one should not expect agents to "tell the truth" regardless of what the others say, but one could hope for agents to "tell the truth" as their best response to the others also "telling the truth." Formally, what this entails is to

replace the game theoretic solution concept from dominance to equilibrium. A Nash equilibrium is a profile of actions, one for each agent, that constitute best responses to one another.[4] Under complete information and using Nash equilibrium, the key necessary condition for implementability of a social choice rule is called monotonicity (defined in Subsection 2.2), and it turns out that, in many relevant classes of environments, monotonicity is compatible with a range of interesting social rules. Some other times, though, monotonicity is still a restrictive condition and then Nash implementability is not possible. For cases like these, it is important to take advantage of the following valuable insight: there is a sense in which monotonicity is an extremely weak condition, if one allows the designer to use random mechanisms and the notion of implementability is slightly relaxed from exact to approximate implementation. Then, it has been shown that essentially every social rule, as long as it satisfies a basic ordinality condition, is approximately implementable in Nash equilibrium.

With incomplete information, the equilibrium notion is Bayesian Nash equilibrium. The key necessary condition for implementability is called incentive compatibility, which says that a social rule must be such that truthful behavior in it is an equilibrium phenomenon. It is easy to see why this becomes a necessary condition: if even when all the others behave truthfully, there is at least one agent who has an incentive to misrepresent his information, the final outcome may be compromised and the desired social rule will not be realized. Incentive compatibility is an important constraint, and in fact, it has become the central condition in the economic theory of information. In addition, a second condition, termed Bayesian monotonicity, is also necessary. This is an extension of the monotonicity condition for incomplete information environments, and it is sometimes extremely restrictive. Fortunately, the approximate approach to implementation is almost as successful as it is with complete information: it turns out that incentive compatibility alone almost always characterizes the set of rules that are approximatley implementable in Bayesian Nash equilibrium.

One should pay tribute to history and, at least, mention briefly the classic early contributions. As antecedents, one must begin with the early

---

[4]The equilibrium paradigm can be understood, at the very least, as the benchmark compatible with the agents' common knowledge of rationality. It can also be seen as the limit of certain dynamic adjustment processes (see Young (1998) for a survey of the learning and evolutionary literatures as possible foundations for equilibrium). Also, in the mechanisms of implementation theory, there is an appealing "focal point" element of their equilibria: namely, "telling the truth," which makes them easy to be "learnable" (see Chen (forthcoming) for supporting experimental evidence.)

4

thinkers in the fields of moral philosophy and political economy, fascinated as they were with the problem of how to reconcile the "private vices" –selfish behavior– and the "public benefits" –socially desirable goals– (e.g., Mandeville (1732), Smith (1776)). Another noteworthy set of precursors came much later. In the first half of the 20th century, socialist and non-socialist writers participated in a debate concerning the advantages of a market system –in terms of decentralized information– versus socialist economies –where all decisions are centrally taken– (Lange (1936, 1937), Hayek (1945)). All these effirts culminated in the fundamental theorems of welfare economics, which relate the performance of decentralized markets to the efficiency of the system (see, for example, Mas-Colell, Whinston and Green (1995, Chapter 16)). In a path-breaking work in mathematical social choice, Arrow (1951) shed light on the problem of aggregating individual preferences into social rankings, and obtained his famous impossibility theorem: there does not exist any non-dictatorial social ranking that is compatible with basic reasonable properties of individual preferences. Arrow's result is a first instance of the difficulties encountered later, for example in the theory of implementation in dominant strategies, although his work did not take agents' incentives into account. Within standard economic theory, the difficulties created to market performance by the presence of public goods were advanced in Samuelson (1954) and by asymmetric information in Akerloff (1970) and Mirrlees (1971). All these contributions prepared the ground for the seminal work of Hurwicz (1960, 1972). Hurwicz's work can be justly called the beginning of the theory of implementation, whose essential components we shall now describe.

2   Preliminaries

We shall consider implementation in the context of a general social choice problem. For simplicity in the presentation, we shall be concerned with finite societies – finite sets of agents– and finite sets of social alternatives.[5]

Let $N = \{1, \ldots, n\}$ be a finite set of *agents* with $n \geq 2$. Let $A = \{a_1, a_2, \ldots, a_k\}$ be a finite set of social *alternatives* or *outcomes*. This set of outcomes is fixed, taken as a primitive, independent of the information held by the agents and not available to the designer.[6]

---

[5]Most of the theory has been developed without the assumption of finite alternatives, but for finite sets of agents; see Mas-Colell and Vives (1993) for results involving a continuum of agents.

[6]If the set of alternatives also depends on the agents' private information, the problem

The private information held by the agents is summarized in the concept of a *state*. The true state will not be verifiable by outsiders (the designeror the court); rather, the agents will have to be given the necessary incentives to reveal it truthfully. We denote by t a typical state and by $\mathcal{T}$ the domain of allowable states. At state t, each agent $i \in N$ is assumed to have a complete and transitive *preference relation* $\succeq_i^t$ over the set A. We denote by $\succ_i^t$ its strict preference part, and by $\sim_i^t$ its indifference relation. We denote by $\succeq^t = (\succeq_1^t, \succeq_2^t, \ldots, \succeq_n^t)$ the profile of preferences in state t. We shall sometimes write $(\succeq_i^t, \succeq_{-i}^t)$ for the profile of preferences in state t, where $\succeq_{-i}^t = (\succeq_1^t, \ldots, \succeq_{i-1}^t, \succeq_{i+1}^t, \ldots, \succeq_n^t)$; the same notational convention will be followed for any profile of objects. Agent i's preferences in state t are represented by a real function, his (Bernoulli) *utility function* $u_i(\cdot, t) : A \times \mathcal{T} \mapsto \mathbb{R}$, i.e., $u_i(a, t) \geq u_i(b, t)$ if and only if $a \succeq_i^t b$.[7]

Fix a state t. We shall refer to the collection $E = \langle N, A, (\succeq_i^t)_{i \in N} \rangle$ as an *environment*. Let $\mathcal{E}$ be the class of allowable environments (since N and A will remain fixed throughout, this class corresponds to all the allowable preference profiles).

At times we shall consider an extension of the model, where lotteries over alternatives are also possible. A *lottery* is a probability distribution over alternatives. Let $\triangle$ denote the set of probability distributions over A. Preferences over lotteries on A are assumed to take the von Neumann-Morgenstern expected utility form (von Neumann and Morgenstern (1944)). That is, abusing notation slightly, given a lottery f in state t, which prescribes alternative a with probability $f(a, t)$, we write $u_i(f, t)$ to refer to agent i's expected utility evaluation of lottery f in state t, i.e., $u_i(f, t) = \sum_{a \in A} f(a, t) u_i(a, t)$.

A *social choice rule* (SCR) F is a mapping $F : \mathcal{E} \mapsto 2^A \setminus \{\emptyset\}$.[8] A *social choice function* (SCF) is a single-valued SCR, and it is denoted by f. When lotteries are permitted, random SCRs and SCFs are defined similarly, where $\triangle$ replaces A in the definition. Again, since N and A are fixed, we shall sometimes abuse notation slightly and write $F(t)$ or $F(\succ^t)$ instead of $F(E)$.

Social choice rules (SCRs) are the objects that the designer would like

---

is more complex because the designer does not know what is really feasible. The classic contribution here is Hurwicz, Maskin and Postlewaite (1995). See also Tian (1989, 1993, 1994), Serrano (1997), Hong (1998), Dagan, Serrano and Volij (1999) for papers that tackle this problem.

[7]The symbol $\mathbb{R}$ denotes the set of real numbers, $\mathbb{R}^l$ the l-dimensional Euclidean space, $\mathbb{R}_+^l$ its non-negative orthant, and $\mathbb{R}_{++}^l$ the interior of this orthant. Also, $\mathbb{Z}$ denotes the set of integer numbers, and $\mathbb{Z}_+$ the set of non-negative integers.

[8]Given two sets S_1 and S_2, $S_1 \setminus S_2$ denotes the set of elements that belong to S_1 and do not belong to S_2. Also, 2^S denotes the power set of S, i.e., the set of all its subsets.

to implement: in each state, she would like to realize some set of outcomes, but unfortunately, she does not know the true state. The implementation problem is precisely when and how this information held by the agents can be elicited so that the desired SCR is successfully implemented. To do this job, the key notion is that of a mechanism.

A *mechanism* $\Gamma = ((M_i)_{i \in N}, g)$ describes a message or strategy set M_i for agent i and an outcome function $g : \prod_{i \in N} M_i \mapsto A$. A random mechanism has the range of the outcome function being $\triangle$. We shall use M_{-i} to express $\prod_{j \neq i} M_j$, and thus, a strategy profile is $m = (m_i, m_{-i})$, where $m_i \in M_i$ and $m_{-i} \in M_{-i}$.

Mechanisms are a representation of the social institution through which the agents interact with the designer and with one another: each agent sends a message to the designer, who chooses an outcome as a function of these strategy choices. We shall only consider static mechanisms in this survey, i.e., those in which each agent sends only one message, and when he does that, he does not know what messages the other agents will be sending.[9] Of course, he can conjecture what those messages might be, and that is the role for the different game theoretic solution concepts that will be employed. Note that a mechanism will in general be played differently by an agent in different states because his preferences over outcomes may change across them. Thus, it makes sense to think of (Γ, E) as the game induced by the mechanism Γ in environment E. We shall confine our attention in this survey to pure strategies: thus, messages are non-random actions.

Let $\mathcal{S}$ be a game theoretic *solution concept*. Given the game (Γ, E), we denote by $\mathcal{S}(\Gamma, E)$ the set of strategy profiles that are recommended by $\mathcal{S}$ in the game (Γ, E). The corresponding set of outcomes will be denoted $g(\mathcal{S}(\Gamma, E))$.

We shall say that the SCR F is $\mathcal{S}$-*implementable* if there exists a mechanism $\Gamma = ((M_i)_{i \in N}, g)$ such that for every $E \in \mathcal{E}$, one has that $g(\mathcal{S}(\Gamma, E)) = F(E)$.

The solution concept $\mathcal{S}$ represents a specific benchmark of rationality that agents obey in a game. Thus, fix $\mathcal{S}$ as a theory of behavior for the agents. Then, two requirements are embodied in the definition of implementability: first, the rational play in the mechanism on the part of the agents produces an outcome prescribed by the desired rule; and second, for each outcome prescribed by the rule, there is a way to implement it when

---

[9]For the very positive results achieved with dynamic mechanisms under complete information, the reader is referred to Abreu and Sen (1990) and Moore and Repullo (1988). See Baliga (1999), Bergin and Sen (1998) and Brusco (1995) for dynamic mechanisms under incomplete information.

agents behave according to the theory of behavior $\mathcal{S}$.[10]

We shall be concerned in the next sections with the different results offered by the theory, as a function of the solution concept employed, and also as a function of the information held by the agents. First, a very appealing solution concept is the idea of dominant strategies. One should aim to construct mechanisms where agents have a dominant strategy, i.e., regardless of what other agents do in the mechanism, each of them always wants to choose his dominant message, and this message does not entail any misrepresentation of his information. Dominance is of course a very strong requirement to describe agents' rational behavior. A less demanding form of rationality in the mechanism is captured by the idea of a Nash equilibrium (Nash (1950a)), in which each agent's message is a best response to the equilibrium messages sent by the others: actions are optimal given expectations, and expectations are justified given the equilibrium actions. Second, one can consider complete information environments, in which the state is common knowledge among the n agents, or incomplete information environments, where this is not the case.[11] The two considerations are related: there is a sense, as we will see, in which dominance is compatible with both types of information.

We shall say that a solution concept is *invariant* if the outcomes it prescribes are independent of equivalent utility representations of the same preferences. Most well-known game theoretic solution concepts (and certainly those based on best responses, such as dominance and equilibrium) are invariant.

2.1 Examples

The following examples give an idea of the range of applications covered by the theory. We present for now the fundamentals of each example, and we shall come back to each of them in the next sections, after going over the corresponding relevant results.

Example 1 *Voting. Let E be an environment, where N is a set of electors (for simplicity, take n to be an odd integer). Let $A = \{a, b\}$ be the set of candidates. Suppose all preferences are strict, i.e., either $a \succ_i^t b$ or $b \succ_i^t a$*

---

[10]This notion is sometimes referred to as full implementation, as opposed to partial or weak implementation, which only asks for the first requirement.

[11]An event is said to be *common knowledge* if everyone knows it, everyone knows that everyone knows it, everyone knows that everyone knows that everyone knows it, and so on (see Aumann (1976)).

for all $i \in N$ and for all $t \in \mathcal{T}$. A state t is a specification of the political preferences of each agent in society (e.g., 80% of electors truly prefer a to b, and 20% b to a). Denote by N_a^t the set of agents that prefer a to b in state t, and let $N_b^t = N \setminus N_a^t$. One standard SCR is majority voting, in which the alternative preferred by the majority is implemented.

Example 2 *A public decision problem. Suppose a community of n agents is interested in undertaking a public project (e.g., the construction of a bridge or a highway, or allocating an amount of public funds to defence). Let $D = \{d_1, \dots, d_k\}$ denote the k different possible public decisions. Agents have preferences over public decisions and money of the quasilinear form: agent i's utility function is $u_i(d, t_i) + x_i$, where we shall interpret u_i as consumer i's valuation of public decision $d \in D$ when t_i is his private information (his true valuation for each decision d is indexed by the variable t_i); the amount x_i is a monetary transfer –tax or subsidy– to agent i. A state t is a profile of the t_i's, indicating the true preferences of the community for each public decision. We take the set A of social alternatives to be*

$$A = \{(d, x_1, \dots, x_n) : d \in D, \quad \forall i, x_i \in \mathbb{R}, \sum_{i \in N} x_i \le 0\}.$$

Thus, although individual subsidies are possible, the project should be funded with the taxes collected for this purpose.

Define an SCF f associated with the decision $d^*(t_1, \dots, t_n)$ for an arbitrary state $t = (t_1, \dots, t_n)$ as that decision that maximizes $\sum_{i \in N} u_i(d, t_i)$, that is, $d^*(t_1, \dots, t_n)$ is the efficient decision if the true state is t because it generates the highest sum of valuations for the public project. The question is whether we can implement this efficient decision.

Example 3 *An exchange economy of private goods. Consider the following environment E, where N is a set of consumers. Consumer i's consumption set is $X_i = \mathbb{R}_+^l$, and each consumer i initially holds an endowment $\omega_i \in X_i$.[12] That is, each agent holds and consumes non-negative amounts of l commodities.[13] Consumer i's wealth is the market value of his initial*

---

[12]In the economic theory of markets, it is useful to separate the consumption and production sides of an economy. In an exchange economy, we abstract from production considerations: the existing goods have already been produced and are in the hands of the consumers. Therefore, the only economic issues concern the reallocation of these goods among consumers through exchange.

[13]To follow most of the literature on exchange economies, we depart in this example from our assumption of a finite set of alternatives. Thus, we allow infinitely divisible commodities.

endowment, i.e., $p \cdot \omega_i$, where p is a vector of market prices. Let the set of alternatives A be the set of allocations or redistributions of the aggregate endowment $\omega = \sum_{i \in N} \omega_i$. That is,

$$A = \{(x_1, \dots, x_n) : \forall i, x_i \in X_i, \quad \sum_{i \in N} x_i \leq \omega\}.$$

A state t is a specification of the true preferences of each consumer. Each agent i is assumed to have a complete and transitive preference relation over the consumption bundles in X_i, i.e., his preferences depend only on his private consumption of goods, and not on that of the other agents: let $x, y \in A$; then, $x \succeq_i^t y$ if and only if $(x_i, z_{-i}) \succeq_i^t (y_i, z'_{-i})$ for any z_{-i}, z'_{-i}, for all $i \in N$ and for all $t \in \mathcal{T}$. In addition, we shall make the standard assumptions that the preference relation $\succeq_i^t$ satisfies:

- *monotonicity: $x_i \succeq_i^t y_i$ whenever $x_{ik} \geq y_{ik}$ for every $k = 1, \dots, l$,*

- *convexity: if $x_i, y_i \succeq_i^t z_i$, $\forall \lambda \in [0, 1]$, $\lambda x_i + (1 - \lambda) y_i \succeq_i^t z_i$, and*

- *continuity: for every pair of sequences $\{x_i^m\}$ and $\{y_i^m\}$, $\lim_{m \to \infty} \{x_i^m\} = x_i^*$ and $\lim_{m \to \infty} \{y_i^m\} = y_i^*$, if for all m $x_i^m \succeq_i^t y_i^m$, then $x_i^* \succeq_i^t y_i^*$.*

The following rule occupies a central position in economics. The Walrasian or competitive SCR assigns to each exchange economy the set of its competitive market equilibrium allocations. That is,

$$F(E) = \{x \in A : \exists p \in \mathbb{R}^l \setminus \{0\} : y \succ_i^t x_i \quad implies \quad p \cdot y > p \cdot \omega_i \forall i \in N\}.$$

An assignment of goods to consumers is a Walrasian or competitive equilibrium allocation if there exist prices such that: (1) each consumer receives in equilibrium a bundle that is top ranked for him among the bundles he can afford, and (2) the aggregate consumption prescribed by the equilibrium allocation is feasible.

Example 4 *King Solomon's dilemma.[14] Two women, Ann and Beth, are claiming to be the mother of a baby. Thus, there are two possible states: t_α in which Ann is the true mother, and t_β, where the mother is Beth. There are four possible outcomes: $A = \{a, b, c, d\}$, where a is to allocate the baby to Ann, b to allocate it to Beth, c is to cut the baby in half, and d is the ominous death-all-around outcome. Ann and Beth's preferences in the two states are as follows:*

$$a \succ_A^\alpha b \succ_A^\alpha c \succ_A^\alpha d; \quad b \succ_B^\alpha c \succ_B^\alpha a \succ_B^\alpha d.$$

---

[14]This example is taken from Glazer and Ma (1989).

$$a \succ_A^\beta c \succ_A^\beta b \succ_A^\beta d; \quad b \succ_B^\beta a \succ_B^\beta c \succ_B^\beta d.$$

Consider the SCF that allocates the baby to the true mother, i.e., $f(t_\alpha) = a$ and $f(t_\beta) = b$. This is the SCF that King Solomon had in mind, but he had a very hard time implementing it. In fact, he cheated.[15] *He proposed a mechanism and then he changed it while it was being played. Namely, he asked the two mothers to report the state, and the outcome he was going to use prescribed a if the unanimous announcement was t_α, b after unanimous report of t_β, and c otherwise. However, when the reports turned out to be non-unanimous, and Ann, the true mother, began to cry in fear of outcome c and changed her announcement to t_β, he implemented a following the unanimous report of t_β. Had Beth known this, she would have probably also started to cry, with the consequent big mess for Solomon.*

2.2   Some Properties of SCRs

We are now interested in describing some important basic properties of SCRs. Ultimately, our aim is the description of those SCRs that are or are not implementable. To begin with, since agents' preferences over alternatives are the important primitive, the theory should be independent of equivalent utility representations of the same preferences. Thus, if one does not allow randomizations, the results will be independent of any monotonically increasing transformation of utility scales. If risk preferences must play a role because of the presence of randomness through lotteries, due to the expected utility hypothesis, the statements will be preserved under any positive affine transformation of utility functions. These considerations lead to the first basic property of SCRs.

In this and the next two sections, we shall always respect the following notational convention: let environment E correspond to state t and preference profile $\succeq^t$, while E' corresponds to t' and $\succeq^{t'}$.

An SCR F is *ordinal* if, whenever $F(E) \neq F(E')$, there must exist an agent $i \in N$ and two alternatives $a, b \in A$ such that $a \succeq_i^t b$ and $b \succ_i^{t'} a$.

That is, for the social choice to change, at least one preference reversal among alternatives must happen. This excludes cardinal rules, in which the social choice may vary with agents' preference intensity, even if no change in the relative ranking of alternatives takes place for any of them.[16]

---

[15] Although the Bible reports this story as proof of his divine wisdom, we ought to question it. Those of us who have had some student cheating on an exam may have the same reaction towards this form of "divine wisdom."

[16] One famous cardinal rule is the utilitarian SCR, where the sum of agents' utilities is maximized.

Proposition 1 *If the SCR F is implementable in any invariant solution concept, it is ordinal.*

Proof: Suppose that $F(E) \neq F(E')$, but that no change in the preference rankings exists between states t and t' corresponding to the two environments. Therefore, $\mathcal{S}(\Gamma, E) = \mathcal{S}(\Gamma, E')$, implying that $g(\mathcal{S}(\Gamma, E)) = g(\mathcal{S}(\Gamma, E'))$, which contradicts that F is implementable. ∎

Apart from ordinality, the following properties will feature prominently in the results in the sequel.

An SCR F is Pareto-efficient or simply *efficient* if for every E and every $a \in F(E)$, there does not exist $b \in A$ such that $b \succ_i^t a$ for every $i \in N$.

Efficiency is a traditional normative desideratum for economics. If an SCR is not efficient, it chooses some alternative that every agent in society ranks below some other fixed alternative. This provides a strong argument against the former alternative as being part of the social choice.

An SCR F is *unanimous* if, whenever $a \succeq_i^t b$ for all $b \in A$ for all $i \in N$, $a \in F(E)$.

Unanimity simply says that if an alternative is top ranked by all individuals, it should be part of the social choice.

An SCR F satisfies *no-veto* if, whenever $a \succeq_i^t b$ for all $b \in A$ and for all individuals i but perhaps one j, then $a \in F(E)$.

This is slightly weaker than unanimity. No-veto says that an alternative will be in the social choice if $(n-1)$ agents rank it at the top of their orderings.

An SCR F is Maskin monotonic or simply *monotonic* if for every $a \in F(E)$, whenever $a \succeq_i^t b$ implies that $a \succeq_i^{t'} b$, one has that $a \in F(E')$.

The logic of this property is the following. Suppose a is in the social choice when preferences are $\succeq^t$. Monotonicity says that, if preferences have changed to $\succeq^{t'}$ in such a way so that no alternative that was indifferent or worse than a has become now better than a, a must remain in the social choice. We shall refer to such a change in preferences as a *monotonic change in preferences around a*. In other words, for a to disappear from the social choice in environment E', it is necessary that at least one individual i has a preference reversal around a in going from E to E' (for him, there was an alternative b such that $a \succeq_i^t b$, but $b \succ_i^{t'} a$).

An SCR F is *dictatorial* if there exists an agent $i \in N$ such that, for every E and every $a \in F(E)$, $a \succeq_i^t b$ for every $b \in A$.

This means that the SCR follows the preferences of a dictator. Most people (presumably, everyone but the dictator) would probably object to this property.

3   Implementation in Dominant Strategies

The first way in which one may want to capture agents' rational behavior in a mechanism is the idea of dominance. We begin by defining the notion of a dominant strategy.

Let $\Gamma = ((M_i)_{i \in N}, g)$ be a mechanism. Message $\hat{m}_i$ is a *dominant strategy* for agent i at state t whenever

$$g(\hat{m}_i, m_{-i}) \succeq_i^t g(m'_i, m_{-i}) \quad \forall m'_i \in M_i, \forall m_{-i} \in M_{-i}.$$

That is, regardless of what the other agents may choose, agent i can never go wrong by choosing m_i if it is dominant. When dominant strategies exist in a game, the prediction of behavior is quite robust; in particular, agent i's behavior is independent of his beliefs about how the others will play the mechanism.[17] It is in this sense, if agent i knows his preferences at state t, that the information he has about the other agents' preferences and his beliefs about how they will play is irrelevant: that is, if he has a dominant strategy, he should choose it, regardless of the other agents' preferences.

Given a mechanism Γ played in state t, we denote by $\mathcal{D}_i(\Gamma, t)$ the set of dominant strategies that agent i has in state t. Let $\mathcal{D}(\Gamma, t) = \prod_{i \in N} \mathcal{D}_i(\Gamma, t)$ be the set of dominant strategy profiles, and $g(\mathcal{D}(\Gamma, t))$ the set of corresponding outcomes.

We shall say that the SCR F is *implementable in dominant strategies* if there exists a mechanism $\Gamma = ((M_i)_{i \in N}, g)$ such that, at every state $t \in \mathcal{T}$, $g(\mathcal{D}(\Gamma, t)) = F(t)$.

Thus, when dominant strategy implementability is possible, it will provide a very robust form of implementation of an SCR. For the rest of the section, we shall concentrate on an important class of specific domains.

In an *independent domain* of preferences: the set of states $calT$ takes the form $\mathcal{T} = \prod_{i \in N} \mathcal{T}_i$, where $\mathcal{T}_i$ is the set of *types* for agent i. Each type $t_i \in \mathcal{T}_i$ is here identified with a preference relation $\succeq_i^{t_i}$ over the set A. Thus, with independent domains, the preference profile at state $t = (t_1, t_2, \ldots, t_n)$ is $\succeq^t = (\succeq_1^{t_1}, \succeq_2^{t_2}, \ldots, \succeq_n^{t_n})$. Note how the assumption of independent domains is simply the requirement that the set of states has the structure of a Cartesian product. As such, at each state, each agent knows his preferences; we will study more general cases in the section on incomplete information.

A direct revelation mechanism, or simply a *direct mechanism*, for SCF f is a mechanism in which the message sets are $M_i = \mathcal{T}_i$ for every $i \in N$ and the outcome function $g = f$.

---

[17]The only small wrinkle to take care of is the possible presence of ties, i.e., agent i may have more than one dominant strategy, and in this case he is indifferent between them.

That is, in a direct mechanism, each agent is simply asked to report his type, anhd the resulting outcome is the one prescribed by f following the reports.

An SCF f is *truthfully implementable in dominant strategies* (or dominant strategy incentive compatible) if for every agent $i \in N$, reporting the truth is a dominant strategy in the direct mechanism for f. An SCR F is truthfully implementable in dominant strategies if every SCF f, single-valued selection of F, is truthfully implementable in dominant strategies.

One important result in the theory of implementation is the *revelation principle*, and we will present it in Section 5. For now, we present a result for implementation in dominant strategies that can be viewed as a stronger version of the revelation principle. This result may explain why the literature on implementation in dominant strategies has for the most part concentrated on direct mechanisms and SCFs (instead of more general mechanisms, and multi-valued SCRs):

Proposition 2 *Assume independent domains and suppose all preference relations are strict. An SCR F is implementable in dominant strategies if and only if it is truthfully implementable in dominant strategies. Moreover, F must be single-valued.*

Proof: Suppose that F is truthfully implementable in dominant strategies, and choose a single-valued selection thereof, which we call g^* with associated direct mechanism Γ^*. Suppose there exist profiles of dominant strategies t and t' in Γ^* when the true state is t. Then, there must be at least one agent i for whom $t_i \neq t_i'$. Since both t_i and t_i' are dominant for agent i, it follows that for every $\hat{t}_{-i}$, $g^*(t_i, \hat{t}_{-i}) = g^*(t_i', \hat{t}_{-i})$. Since this holds for every agent i for whom $t_i \neq t_i'$, it follows that $g^*(t) = g^*(t')$. Thus, if there are multiple dominant strategy profiles, their outcomes must be the same. This last argument also demonstrates that F must be single-valued, and therefore, it is implementable in dominant strategies.

Suppose now that F is implementable in dominant strategies. That is, there exists a mechanism Γ such that for every state t, each agent i has a dominant strategy $m_i(t)$ satisfying that $g(m(t)) \in F(t)$. By independent domains, one can write that $m_i(t) = m_i(t_i)$ because agent i's preferences do not depend on t_{-i}.

Consider the direct mechanism Γ^*, where g^* is a single-valued selection of F, $g^*(t_1, \ldots, t_n) = g(m_1(t_1), \ldots, m_n(t_n))$. Since for each $i \in N$, $m_i(t_i)$ is dominant in Γ, one has that:

$$g(m_i(t_i), m_{-i}) \succeq_i^{t_i} g(m_i, m_{-i}) \qquad \forall m_i \in M_i, \forall m_{-i} \in M_{-i},$$

where, by strict preferences, the indifference happens only when $g(m_i(t_i), m_{-i}) = g(m_i, m_{-i})$. By a similar argument as above, if an agent has more than one dominant strategy in Γ, they must prescribe the same outcomes. Since this is true for each such agent, it follows that F must be single-valued.

In addition, the above preference comparison holds for the m_is and the m_{-i}s that are dominant for some t_i and t_{-i}. It follows that declaring the truth is also dominant in the direct mechanism Γ^*. This shows that F is truthfully implementable in dominant strategies.

$\blacksquare$

Laffont and Maskin (1982) present other conditions under which this result holds. However, when indifferences are allowed, even under independent domains, truthful implementation of an SCF in dominant strategies does not imply its full implementation, nor that it is a singleton. Examples are easy to construct. Outside of independent domains, truthful implementability in dominant strategies is also a necessary condition for full implementability. This last fact justifies that we now focus on truthful implementability in dominant strategies of SCFs. A necessary condition for this is strategy-proofness.

An SCF f is *strategy-proof* if for every $i \in N$, for every $t_{-i} \in \mathcal{T}_{-i}$ and for every $t_i, t_i' \in \mathcal{T}_i$, one has that

$$f(t_i, t_{-i}) \succeq_i^t f(t_i', t_{-i}).$$

Strategy-proofness means that, no matter what the value of t_{-i} is – information held by the others about their preferences–, no agent would benefit from misrepresenting his own information. It is easy to see how strategy-proofness is necessary for truthful implementation in dominant strategies. With independent domains, it is also sufficient.[18]

If the allowable domain is rich enough, most rules are not strategy-proof. This is the content of the next pages, which will arrive at an important conclusion, that provided by the Gibbard-Satterthwaite theorem (proved independently by Gibbard (1973) and Satterthwaite (1975)).

We shall present a proof of the Gibbard-Satterthwaite theorem that combines elements of Reny (2001) and Mas-Colell, Whinston and Green (1995).[19] We begin by showing a result that is stronger than the Gibbard-

---

[18]Beyond independent domains, the difficulty is the insufficiency of the direct revelation mechanism with strategy sets T_i.

[19]Reny's proof follows closely the elegant proofs of Arrow's impossibility theorem of Geanakoplos (1996). See also alternative proofs due to Barberá (1983) and Benoit (1999). The proof that we present will be consistent with our maintained assumption of a finite set of alternatives. See Barberá and Peleg (1990) for the more general case.

Satterthwaite theorem. By relying on monotonicity, it will connect nicely with results in the next sections.

Consider the *unrestricted domain* of strict preference profiles; call it $\mathcal{T}^S$, i.e.,

$$\mathcal{T}^S = \{(t_1, \ldots, t_n) \in calT : \forall i \in N, a \sim_i^{t_i} b \quad \text{if and only if} \quad a = b\}.$$

Thus, we begin by exploring preference profiles in which there are no indifferences. This will be the case for our next two results. Later we shall consider also preference profiles with indifferences.

Proposition 3 *Suppose the allowable domain is $\mathcal{T}^S$. If $|A| \geq 3$ and $f : \mathcal{T}^S \mapsto A$ is a unanimous and monotonic SCF, it is dictatorial.*

Proof:

Step 1: Consider any two alternatives $a, b \in A$ and a preference profile such that a is top ranked for every agent and b is bottom ranked for every agent. By unanimity, the social choice is a. Consider agent 1's ranking and, in it, raise b one position at a time. By monotonicity, as long as b does not rise over a the social choice remains a. When b jumps over a in 1's ranking, the only possible social choices are either a or b, also by monotonicity. If it remains being a, begin the same process with agent 2, and so on. The point is that there must exist one agent for whom when a falls below b in his ranking, the social choice switches to b: by unanimity, otherwise we would end up with a profile where b is top ranked for all agents and is not the social choice, a contradiction. Call this individual j.

To be clear, refer to the profile right before preference between a and b change for agent j as profile t^1, i.e.,

$$b \succ_1^1 a \succ_1^1 \ldots$$
$$\ldots$$
$$b \succ_{j-1}^1 a \succ_{j-1}^1 \ldots$$
$$a \succ_j^1 b \succ_j^1 \ldots$$
$$a \succ_{j+1}^1 \ldots \succ_{j+1}^1 b$$
$$\ldots$$
$$a \succ_n^1 \ldots \succ_n^1 b$$

yielding that $f(\succ^1) = a$.

Call the profile that results from switching a and b in agent j's preferences profile t^2:

$$b \succ_1^2 a \succ_1^2 \dots$$

$$\dots$$

$$b \succ_{j-1}^2 a \succ_{j-1}^2 \dots$$
$$b \succ_j^2 a \succ_j^2 \dots$$
$$a \succ_{j+1}^2 \dots \succ_{j+1}^2 b$$

$$\dots$$

$$a \succ_n^2 \dots \succ_n^2 b$$

yielding that $f(\succ^2) = b$.

Step 2: Next consider the profiles $t^{1'}$ and $t^{2'}$ as the following variants of profiles t^1 and t^2, respectively: for agents $i < j$, send a to the bottom of their rankings, and for agents $i > j$ send a to the second place from the bottom of their rankings. That is,

$$b \succ_1^{1'} \dots \succ_1^{1'} a$$

$$\dots$$

$$b \succ_{j-1}^{1'} \dots \succ_{j-1}^{1'} a$$
$$a \succ_j^{1'} b \succ_j^{1'} \dots$$
$$\dots \succ_{j+1}^{1'} a \succ_{j+1}^{1'} b$$

$$\dots$$

$$\dots \succ_n^{1'} a \succ_n^{1'} b$$

and

$$b \succ_1^{2'} \dots \succ_1^{2'} a$$

$$\dots$$

$$b \succ_{j-1}^{2'} \dots \succ_{j-1}^{2'} a$$
$$b \succ_j^{2'} a \succ_j^{2'} \dots$$
$$\dots \succ_{j+1}^{2'} a \succ_{j+1}^{2'} b$$

$$\dots$$

$$\dots \succ_n^{2'} a \succ_n^{2'} b.$$

We claim that $f(\succ^{1'}) = a$ and $f(\succ^{2'}) = b$. First, recall that $f(\succ^2) = b$. In going from $\succ^2$ to $\succ^{2'}$, no change in preferences around b has taken place. Therefore, since f is monotonic, $f(\succ^{2'}) = b$. Next, note that in going from $\succ^{2'}$ to $\succ^{1'}$, the only change takes place in the ranking of alternatives a and b for agent j. Monotonicity therefore implies that $f(\succ^{1'})$ must be either a or b. However, it cannot be b: monotonicity then would imply that $f(\succ^1) = b$, which is a contradiction. Therefore, as claimed, $f(\succ^{1'}) = a$.

Step 3: Let c be a third alternative, distinct from a and b. Consider the following preference profile (t^3) obtained from $t^{1'}$ in a way that results in a not changing rankings with any alternative for any individual:

$$\ldots \succ_1^3 c \succ_1^3 b \succ_1^3 a$$
$$\ldots$$
$$\ldots \succ_{j-1}^3 c \succ_{j-1}^3 b \succ_{j-1}^3 a$$
$$a \succ_j^3 c \succ_j^3 b \succ_j^3 \ldots$$
$$\ldots \succ_{j+1}^3 c \succ_{j+1}^3 a \succ_{j+1}^3 b$$
$$\ldots$$
$$\ldots \succ_n^3 c \succ_n^3 a \succ_n^3 b.$$

Since $f(\succ^{1'}) = a$, monotonicity implies that $f(\succ^3) = a$.

Step 4: Next, from profile t^3, just switch preferences between a and b for agents $i > j$ to get profile t^4. That is,

$$\ldots \succ_1^4 c \succ_1^4 b \succ_1^4 a$$
$$\ldots$$
$$\ldots \succ_{j-1}^4 c \succ_{j-1}^4 b \succ_{j-1}^4 a$$
$$a \succ_j^4 c \succ_j^4 b \succ_j^4 \ldots$$
$$\ldots \succ_{j+1}^4 c \succ_{j+1}^4 b \succ_{j+1}^4 a$$
$$\ldots$$
$$\ldots \succ_n^4 c \succ_n^4 b \succ_n^4 a.$$

Since in going from $\succ^3$ to $\succ^4$, preferences involving alternatives other than a and b are unaffected, monotonicity implies that $f(\succ^4)$ is either a or b. However, it cannot be b: if it were and we had a change in preferences so that c becomes top-ranked for every individual, monotonicity would imply that the social choice then would continue to be b, causing an eventual violation of unanimity. Therefore, $f(\succ^4) = a$.

Step 5: Note that any profile where a is top ranked for agent j can be generated from profile t^4 in a monotonic way with respect to a. Therefore, monotonicity implies that for any such profile, the social choice remains a.

Thus, for alternative a, agent j is a dictator over states that involve only strict preferences. Since a was chosen arbitrarily, this establishes that for each alternative, one can find a dictator over strict preference profiles. But one cannot have more than one dictator: suppose agent j is the dictator for alternative a and agent j' for alternative a'. Consider a strict preference profile $\succ$ in which alternative a is top ranked for agent j and a' is top ranked for agent j'. Because agent j is a dictator for a, one has that $f(\succ) = a$, and since j' is a dictator for a', one has that $f(\succ) = a'$, a contradiction.

Therefore, there is only one dictator for every alternative over all preference profiles in $\mathcal{T}^S$.

∎

Next, we relate the assumptions of the previous theorem to strategy-proofness (a version of this result appears in Dasgupta, Hammond and Maskin (1979)).

Proposition 4 *If $f : \mathcal{T}^S \mapsto A$ is strategy-proof and onto, then it is unanimous and monotonic.*

Proof: Let $\succ$ be a profile such that $f(\succ) = a$ and consider a monotonic change in preferences around a to get to profile $\succ'$. That is, for every agent i, $a \succ_i b$ implies that $a \succ'_i b$. We claim that $f(\succ') = a$, so that f is monotonic.

To see this, we first show that $f(\succ'_1, \succ_{-1}) = a$. By strategy-proofness, $f(\succ) \succ_1 f(\succ'_1, \succ_{-1})$ if $f(\succ) \neq f(\succ'_1, \succ_{-1})$. Since the preference change for agent 1 is monotonic around $f(\succ)$, one gets that $f(\succ) \succ'_1 f(\succ'_1, \succ_{-1})$, which contradicts strategy-proofness at the profile $\succ'$. Hence, as claimed, $f(\succ'_1, \succ_{-1}) = a$.

Since we can get from profile $\succ$ to profile $\succ'$ by changing agents' preferences one at a time in the same way, it follows that $f(\succ') = a$, as we wanted to show. Thus, f is monotonic.

To show unanimity, choose $a \in A$. Since f is onto, there exists $\succ$ such that $f(\succ) = a$. By monotonicity, $f(\succ') = a$ if $\succ'$ is obtained from $\succ$ simply by pushing a to the top of every agent's ranking. Monotonicity again implies that for any preference profile $\succ''$ where a is top-ranked for each individual, $f(\succ'') = a$, which shows unanimity.

∎

Our next step is to also consider profiles with indifferences, i.e., states outside of the set $\mathcal{T}^S$. Indeed, these two previous propositions are important in establishing the Gibbard-Satterthwaite theorem:

Proposition 5 *If $|A| \geq 3$, the domain of preferences includes all possible strict rankings, and f is strategy-proof and onto, f is dictatorial.*

Proof: The previous two propositions already imply the result over the domain $\mathcal{T}^S$. Thus, consider a preference profile $\succeq$ in which indifferences occur.

We argue by contradiction, i.e., suppose f is not dictatorial. Call j the agent that is a dictator for f over all profiles in $\mathcal{T}^S$, assume that $f(\succeq) = a$ and there exists $b \in A$ such that $b \succ_j^{t_j} a$. Let b be top-ranked for agent j under preferences $\succeq_j^{t_j}$.

Consider now another preference profile $\succeq'$ such that:

- (i) $\succeq'_i$ is a strict preference relation for all $i \in N$;

- (ii) for all $i \neq j$, $a \succ'_i b \succ'_i c$ for every $c \neq a, b$; and

- (iii) $b \succ'_j a \succ'_j c$ for every $c \neq a, b$.

Let $k \neq j$ and consider the profile $(\succeq'_k, \succeq_{-k})$. By strategy-proofness applied at this profile, we must have that $f(\succeq'_k, \succeq_{-k}) = a = f(\succeq)$. With the same logic, and moving agents other than j one by one from preference $\succeq_i$ to $\succeq'_i$, we conclude that $f(\succeq_j, \succeq'_{-j}) = a = f(\succeq)$.

Next, note that strategy-proofness at profile $\succeq'$ implies that $f(\succeq')$ is either a or b. However, it cannot be b, by strategy-proofness applied to agent j at profile $(\succeq_j, \succeq'_{-j})$. Thus, $f(\succeq') = a$, which contradicts that j is a dictator over strict profiles. We conclude that f is dictatorial.
∎

The unrestricted domain of strict preferences assumption is strong because every possible strict ordering of alternatives must be allowed. On the other hand, there are some interesting subdomains to which the impossibility result extends. Remarkably, though, there are also important subdomains in which possibility results arise, such as that of quasilinear preferences (Example 2).[20] Likewise, the assumption of having at least three alternatives also matters, as we are about to see in Example 5. Finally, versions of the impossibility result have been obtained for SCRs that are not single-valued and that use lotteries over alternatives.[21]

Example 5 *Recall Example 1. Consider the majoritarian SCF f: $f(E) = a$ whenever $|N_a^t| > |N_b^t|$, and $f(E) = b$ otherwise. Suppose we use a direct revelation mechanism, in which each agent is simply asked to report his preferences (say, by reporting his top ranked alternative), and the outcome function implements the majoritarian rule on the basis of the collected reports. In the above notation, let $M_i = \{a, b\}$ for every $i \in N$ be the message set, and $g(m) = f(m)$. It is easy to see that reporting the truth is a dominant strategy in this mechanism, i.e., regardless of what the other agents report, no agent has an incentive to lie about his preferences. To see this, note that agent i's report changes the outcome only when it is pivotal, i.e.,*

---

[20]Other restrictions, such as separability (Barberá, Sonnenschein and Zhou (1991)) or single-peakedness of preferences (Sprumont (1991)) also yield very interesting positive results.

[21]See Barberá (1977), Barberá, Bogomolnia and van der Stel (1998), Barberá, Dutta and Sen (2001), Benoit (2002), Ching and Zhou (2002), and Duggan and Schwarz (2000).

when there is a tie between the two alternatives in the reports of the other agents, and then he prefers to tell the truth. Thus, f is implementable in dominant strategies, but it is not dictatorial.

Example 6 *Recall the efficient decision of Example 2. Consider a direct revelation mechanism, in which every agent is asked his type, agent i then announces m_i and then $d^*(m)$ is undertaken and certain transfers $x_i(m)$ are imposed.*

The Vickrey-Clarke-Groves (Vickrey (1961), Clarke (1971), Groves (1973)) mechanism implements the efficient decision in dominant strategies.[22] That is, regardless of the other agents' announcements, agent i will always want to tell the truth. This is done by setting transfers $x_i(m) = \sum_{j \neq i} u_j(d^*(m), m_j) + h_i(m_{-i})$, where we choose the functions $h_i(\cdot)$ to ensure that $\sum_{i \in N} x_i(m) = 0$. Now we show that every agent $i \in N$ maximizes his utility at $m_i = t_i$ regardless of the other agents' announcements. Suppose not, that is, there exist announcements m_{-i} for the other agents such that agent i of type t_i prefers to announce τ_i rather than t_i. Namely,

$$u_i(d^*(\tau_i, m_{-i}), t_i) + \sum_{j \neq i} u_j(d^*(\tau_i, m_{-i}), m_j) > u_i(d^*(t_i, m_{-i}), t_i) + \sum_{j \neq i} u_j(d^*(t_i, m_{-i}), m_j),$$

but this inequality contradicts the definition of d^* for the state (t_i, m_{-i}). Therefore, this mechanism induces truth telling as a dominant strategy.[23]

4 Implementation in Nash Equilibrium: Complete Information

One important message delivered by the Gibbard-Satterthwaite theorem of the previous section is that implementation in dominant strategies has very limited success over the unrestricted domain of preferences when there are at least three alternatives, since only dictatorial rules can be implemented. One possible way out of this strongly negative result, as Example 6 demonstrated, is to restrict the domain of allowable preferences to some relevant subdomain. The other way out is the one that we explore in this section,

---

[22]Vicrey's work deals with the allocation of a private good for money through an auction procedure. The reader can see that this can be another application of the model in this example.

[23]There are difficulties concerning balanced-budget transfers for general functions $u_i(\cdot)$; see D'Aspremont and Gerard-Varet (1979) and Green and Laffont (1979) for results tackling this problem.

and it concerns a change in the theory of behavior that we will have for the agents when confronting a mechanism. Instead of dominance, the new idea will be that of an equilibrium.

We assume that there is *complete information* among the agents, i.e., all n agents know the true state t, all know that they know it, and so on. This assumption is especially justified when the implementation problem concerns a small number of agents that hold good information about one another (think, for example, of a setting in which two parties are trying to write a contract, and that they know something about each other that would be very difficult to verify for an outside enforcer). We also drop the assumption of independent domains of preferences.

Given a mechanism $\Gamma = ((M_i)_{i \in N}, g)$ played in state t, a *Nash Equilibrium* of the mechanism Γ in state t is a strategy profile m^* such that:

$$\forall i \in N, g(m^*(t)) \succeq_i^t g(m_i, m^*_{-i}(t)) \forall m_i \in M_i.$$

At a Nash equilibrium, each agent is choosing a strategy that is a best response to the equilibrium strategies employed by the others. Note the "fixed point" idea between actions and expectations embodied in the concept: each agent i chooses m_i^* because he expects the others to play m^*_{-i}, and these expectations are justified because each of the others have no incentive to choose something else if they all stick to m^*. Thus, while dominance requires an agent to choose his strategy regardless of what the others play, the equilibrium logic asks an agent to choose his strategy as a best response to what he conjectures the others will be doing.

Given a mechanism $\Gamma = ((M_i)_{i \in N}, g)$ played in state t, we denote by $\mathcal{N}(\Gamma, t)$ the set of Nash equilibria of the game induced by Γ in state t. Likewise, $g(\mathcal{N}(\Gamma, t))$ denotes the corresponding set of Nash equilibrium outcomes.

We shall say that an SCR F is *Nash implementable* if there exists a mechanism $\Gamma = ((M_i)_{i \in N}, g)$ such that for every $t \in \mathcal{T}$, $g(\mathcal{N}(\Gamma, t)) = F(t)$.

A first observation is in order. To achieve (full) implementability, in general we will have to go well beyond direct mechanisms. Assume the existence of a universally bad alternative (for example, an allocation that gives zero amounts of all goods to every consumer in an exchange economy). Given the complete information assumption, a direct mechanism for an SCF f would ask each agent to report the state. Let us then use a simple direct mechanism, in which if all agents agree on the announced state $\hat{t}$, $f(\hat{t})$ is implemented, while the bad alternative results otherwise. Note how, in addition to the "good" Nash equilibrium of the mechanism, in which all agents

report the true state, $\hat{t} = t$, there are multiple equilibrium outcomes corresponding to any other unanimous announcement. It follows that, to achieve our notion of implementability, we will have to employ more sophisticated mechanisms.

We shall now be concerned with the investigation of necessary and sufficient conditions on SCRs that are Nash implementable. The next fundamental result is due to Maskin (1999).[24]

Proposition 6 *If an SCR F is Nash implementable, it is monotonic.*

Proof: Suppose the true state is t, and let $a \in F(t)$. Because F is Nash implementable, there exists a mechanism Γ and a Nash equilibrium thereof m^* played in state t such that $g(m^*) = a$. Now consider a monotonic change in preferences around a to get to state t'. Since no alternative has risen in any agent' preference ranking with respect to a, the profile m^* continues to be a Nash equilibrium of Γ in state t'. That is, $a \in g(\mathcal{N}(\Gamma, t'))$ and, since F is Nash implementable, $a \in F(t')$, but then F is monotonic. ■

Thus, monotonicity is a necessary condition for Nash implementability in any environment. The next question we may want to ask is how restrictive monotonicity is. When applied to SCFs on the unrestricted domain of preferences, monotonicity is a very demanding condition. Recall Proposition 3, which establishes that with at least three alternatives monotonicity and unanimity imply that the SCF is dictatorial.[25] Note that Proposition 3 can be extended to domains that allow indifferences: indeed, one can add a sixth step to the existing proof, in which one can undo indifferences in a monotonic way around the alternative chosen by the SCF to end up showing that there exists a dictator. Maintaining the unrestricted domain assumption, monotonicity is still a very strong requirement for SCRs: Hurwicz and Schmeidler (1978) show that monotonicity is incompatible with efficiency.

However, for multi-valued SCRs defined over many interesting restricted domains, monotonicity is more permissive. For example, one can show that the Walrasian SCR of Example 3 is monotonic in certain domains (see Example 7 below). Also, the SCR that assigns to each social choice problem the set of its efficient alternatives is monotonic in any environment. In general,

---

[24]Although the first version of his results for Nash implementation was circulated in an MIT working paper in 1977, Maskin could not implement their publication until 22 years later.

[25] See Mueller and Satterthwaite (1977), Dasgupta, Hammond and Maskin (1979) and Saijo (1987).

we can conclude that monotonicity is compatible with a range of interesting social goals in relevant domains.

Our next step is to inquire about the sufficient conditions for Nash implementability. As it turns out, Proposition 6 has almost a converse, at least for the case of three or more agents. The next result is also due to Maskin (1999).[26]

Proposition 7 *Let $n \geq 3$. If an SCR F is monotonic and satisfies no-veto, it is Nash implementable.*

Proof: The proof is based on the construction of a canonical mechanism that will Nash implement F under the two conditions assumed.

Consider the following mechanism $\Gamma = ((M_i)_{i \in N}, g)$, where agent i's message set is $M_i = A \times \mathcal{T} \times \mathbb{Z}_+$ (recall that A is the set of alternatives, $\mathcal{T}$ the set of allowable states, and $\mathbb{Z}_+$ the set of non-negative integers). We shall denote a typical message sent by agent i by $m_i = (a^i, t^i, z^i)$. The outcome function g is defined in the following three rules:

- (i) If for every agent $i \in N$, $m_i = (a, t, 0)$ and $a \in F(t)$, $g(m) = a$.

- (ii) If $(n-1)$ agents $i \neq j$ send $m_i = (a, t, 0)$ and $a \in F(t)$, but agent j sends $m_j = (a^j, t^j, z^j) \neq (a, t, 0)$, then $g(m) = a$ if $a^j \succ_j^t a$, and $g(m) = a^j$ otherwise.

- (iii) In all other cases, $g(m) = a'$, where a' is the alternative chosen by the agent with the lowest index among those who announce the highest integer.

We now have to show that for all $t \in \mathcal{T}$, the set of Nash equilibrium outcomes of the mechanism Γ coincides with $F(t)$, i.e., $g(\mathcal{N}(\Gamma, t)) = F(t)$. *Step 1:* For all $t \in \mathcal{T}$, $F(t) \subseteq g(\mathcal{N}(\Gamma, t))$. Fix state $t \in \mathcal{T}$. Let $a \in F(t)$ and consider the following strategy profile used by the agents, where each agent $i \in N$ chooses $m_i^* = (a, t, 0)$. First, note that this profile falls under Rule (i) of the outcome function and a would be implemented. Furthermore, no agent i has an incentive to deviate from m_i^*: by deviating, he could only hope to induce Rule (ii) (because Rule (iii) is not accessible with a unilateral deviation from this strategy profile). But since the strategy profile m^* includes a unanimous report of the true state, agent i could only change

---

[26]We present a proof due to Repullo (1987); see alternative proofs in Williams (1986) and Saijo (1988).

the outcome if he chose an alternative that he does not prefer to a. Thus, there is no unilateral profitable deviation and m^* is a Nash equilibrium.

Step 2: For all $t \in \mathcal{T}$, $g(\mathcal{N}(\Gamma, t)) \subseteq F(t)$. Fix a state $t \in \mathcal{T}$. Let $\hat{m} \in \mathcal{N}(\Gamma, t)$ and let $\hat{a}$ be the corresponding outcome according to the outcome function g. If $\hat{a}$ is a result of either Rule (ii) or Rule (iii), there exists $j \in N$ such that every $k \neq j$ can induce his top ranked outcome by choosing a high enough integer. Therefore, $\hat{a}$ must be top ranked for at least $(n-1)$ agents, and by no-veto, $\hat{a} \in F(t)$.

We are left with $\hat{a}$ being a result of Rule (i). That is, there is a unanimous report $\hat{m} = (\hat{a}, \hat{t}, 0)$ with $\hat{a} \in F(\hat{t})$, but $\hat{t} \neq t$, i.e., a false state is being reported. If $\hat{a} \in F(t)$, we are done. So suppose that $\hat{a} \notin F(t)$. That is, $\hat{a} \in F(\hat{t})$ and $\hat{a} \notin F(t)$. Since F is monotonic, in going from state $\hat{t}$ to state t, a preference reversal around $\hat{a}$ must have occurred, i.e., there exists $i \in N$ and $b \in A$ such that $\hat{a} \succeq_i^{\hat{t}} b$ and $b \succ_i^t \hat{a}$. Then, consider the following deviation on the part of agent i from $\hat{m}$: let agent i send message $m_i' = (b, \cdot, \cdot)$. By doing this, outcome b is implemented and agent i profits from the deviation (recall that his true preferences are the ones described in profile t), thereby contradicting that $\hat{m}$ is a Nash equilibrium. Thus, this case is impossible and the proof is complete.

∎

In closing the gap between necessary and sufficient conditions, it is first important to note the following domain of environments:

An environment is *economic* if, as part of the social alternatives, there exists a private good –e.g., money– over which all agents have a strictly positive preference.

Note then that in economic environments the no-veto condition is vacuously satisfied, because it is never the case that an alternative is top ranked by $(n-1)$ individuals. We obtain the following corollary:

Corollary 1 *Consider economic environments and let $n \geq 3$. An SCR F is Nash implementable if and only if it is monotonic.*

In general, though, no-veto is not a necessary condition for Nash implementability, and monotonicity alone is not sufficient (see Maskin (1999)).[27] Necessary and sufficient conditions were provided for general environments in Moore and Repullo (1990).

The two-agent case is somewhat special. Recall Rule (ii) in the canonical mechanism of the proof of Proposition 7. With more than two agents, it

---

[27] Although there are other environments where monotonicity alone is enough to guarantee Nash implementability (e.g., Moulin (1983)).

is possible to detect unilateral deviations from an otherwise unanimous announcement, while this is not true if one considers two-agent environments. Nonetheless, this difficulty can be circumvented, and necessary and sufficient conditions were provided in Moore and Repullo (1990) and Dutta and Sen (1991).

One final observation is worth making. The canonical mechanism in the proof of Proposition 7 is necessarily quite abstract, as it is able to handle a large number of social choice problems. In particular, the device in Rule (iii), called an integer game, has received much criticism for being unnatural.[28] However, one should not lose track of the main purpose of the current exercise, and that is the characterization of rules that are Nash implementable. As such, integer games are just a method of proof employed in establishing this important result. It should be true, though, and indeed this is the case, that more realistic mechanisms can be constructed when one deals with a specific application of the theory.

We now revisit some of our examples.

Example 7 *Recall the exchange economy of Example 3. It is easy to see that the Walrasian rule is manipulable (not strategy-proof), in the sense that agents may have an incentive to misrepresent their true preferences. For example, consider a specific domain consisting of two economies. In both, the set of agents is $N = \{i, j\}$, there are two consumption goods 1 and 2, and the initial endowment is $\omega_i = (3, 9)$ and $\omega_j = (9, 3)$. Suppose there is no uncertainty about the preferences of agent i, which are represented by the utility function $u_i(x_{i1}, x_{i2}) = x_{i1}x_{i2}$. However, while in state t, agent j's utility function is $u_j((x_{j1}, x_{j2}), t) = x_{j1}x_{j2}$, in state t', it is $u_j((x_{j1}, x_{j2}), t') = x_{j1}^2 x_{j2}$. The reader can check that the unique competitive allocation in state t assigns the bundle $(6, 6)$ to each consumer (supported by prices $p^t = (1, 1)$), while in state t' it is the allocation $((60/13, 20/3), (96/13, 16/3))$ (supported by prices $p^{t'} = (13/9, 1)$). Note how $u_j((96/13, 16/3), t) > u_j((6, 6), t)$, and therefore, if asked by the planner about his preferences before the Walrasian rule is implemented, agent j has an incentive to lie.*

One possible way out for the designer is to take advantage of the fact that the state is common knowledge among the agents (complete information). The hope is that the truth can be elicited, if not as a dominant strategy (as we have just established), perhaps as the Nash equilibrium of some mechanism. The key is to show that, in the specified environments, the Walrasian

---

[28]The lack of compactness of the strategy set implied by the use of $\mathbb{Z}_+$ is not a great concern: it can be remedied by introducing a modulo game (a compactified version of the integer game); see Jackson (1992) for an insightful critique.

SCR is monotonic. This is indeed the case if one considers economies in which the Walrasian allocations assign positive amounts of all goods to each consumer, as the reader can check. Difficulties with monotonicity appear at the boundary of the consumption set (see Hurwicz, Maskin and Postlewaite (1995)).

There are interesting domains of economies where the Walrasian SCR assigns strictly positive bundles to each consumer. For example, suppose that all l goods considered in the model are essential to subsistence (all bundles along the axes of the non-negative orthant are strictly less preferred than those in its interior) and assume that each consumer holds initially positive amounts of all goods. Then, it follows from Corollary 1 that the Walrasian SCR is Nash implementable in this domain of allowable environments when there are at least three consumers.[29]

Example 8 *Recall Example 4 and the Solomonic rule prescribed there. It turns out that this SCF is not easy to implement. First, it follows from our discussion in Example 4 that it is not implementable in dominant strategies. Moreover, it is not implementable in Nash equilibrium either, because it is not monotonic (recall Proposition 6). That is, note that the social choice changes from state t_α to state t_β, even though no alternative that was less preferred than a for either woman has become now preferred to a. Therefore, there does not exist any mechanism that can implement the Solomonic rule in Nash equilibrium.*

4.1   Virtual Implementation in Nash Equilibrium

Thus far we have understood that monotonicity is the key condition behind Nash implementability, and we have seen that in certain environments (e.g., as in Example 7), it is a permissive condition in that it is compatible with interesting social goals. However, as pointed out in Example 8 (recall also the work cited in footnote 25), some other times it imposes severe restrictions on the set of Nash implementable rules.

In this subsection, two important changes are introduced in the model. First, the set of alternatives considered is the set $\triangle$ of lotteries over the set A, and second, the designer will content herself with approximately implementing her SCR, instead of implementing it exactly.[30]   One can easily

---

[29]Indeed, apart from the canonical mechanism of the proof of Proposition 7, mechanisms tailored to the Walrasian SCR for private and public goods have been proposed in Hurwicz (1979), Schmeidler (1980), Walker (1981), Dutta, Sen and Vohra (1995).

[30]Although I prefer the more accurate name of "approximate implementation," the

justify both changes. On the one hand, lotteries are devices used often in the allocation of resources. On the other hand, the approximation to the desired SCR will be an arbitrarily close one, i.e., for all $\epsilon \in (0,1)$, the desired SCR will be implemented with probability $(1 - \epsilon)$, while something else will be allowed to happen with probability ϵ. The effects of these two changes on the scope of the theory of implementation will be striking. This is an amazing insight, first obtained independently by Matsushima (1988) and Abreu and Sen (1991).

Let us now recall some of our definitions, properly adapted to the consideration of lotteries.

Let $A = \{a_1, a_2, \ldots, a_k\}$ be the finite set of social alternatives. Let $\triangle$ be the set of lotteries (probability distributions) over the set A, i.e.,

$$\triangle = \{(p_1, p_2, \ldots, p_k) \in \mathbb{R}_+^k : \sum_{j=1}^{k} p_j = 1\}.$$

An SCR F is now a non-empty valued mapping $F : \mathcal{T} \mapsto 2^{\triangle} \setminus \{\emptyset\}$. Of course, non-random alternatives and non-random SCRs are covered as particular cases of these definitions.

Let $\triangle_+$ be the subset of $\triangle$ of strictly positive lotteries, i.e.,

$$\triangle_+ = \{(p_1, p_2, \ldots, p_k) \in \mathbb{R}_{++}^k : \sum_{j=1}^{k} p_j = 1\}.$$

As we shall see shortly, the set $\triangle_+$ will play an important role in the analysis.

In order to speak of virtual implementation, a notion of "closeness" is called for. Given two lotteries $p, p' \in \triangle$, $d(p, p')$ will refer to the Euclidean distance between them, i.e.,

$$d(p, p') = [\sum_{j=1}^{k} (p_j - p_j')^2]^{1/2}.$$

Given two SCRs F and H, we define the following distance between them at state t. For every state $t \in \mathcal{T}$, let π_t be a bijection, $\pi_t : F(t) \mapsto H(t)$. Then,
$$d(F(t), H(t)) = \sup_{p \in F(t)} d(p, \pi_t(p)).$$

literature has referred to this approach as virtual implementation.

28

Recall also that in the mechanism $\Gamma = ((M_i)_{i \in N}, g)$, the outcome function $g : \prod_{i \in N} M_i \mapsto \triangle$, and that $g(\mathcal{N}(\Gamma, t))$ denotes the set of Nash equilibrium outcomes of the mechanism Γ in state t.

We shall say that an SCR F is *virtually Nash implementable* if for every $\epsilon > 0$, there exists a mechanism Γ such that for every state $t \in \mathcal{T}$, $d(F(t), g(\mathcal{N}(\Gamma, t))) < \epsilon$.

Note how (exact) Nash implementability corresponds to this definition, when $\epsilon = 0$. virtual implementability allows the designer to "make a mistake" with respect to her true goals, but that "mistake" can be made arbitrarily small. Stated in slightly different terms, virtual implementability amounts to exact implementability of a near-by rule.

Following our line of inquiry, we are interested in identifying the necessary and sufficient conditions for virtual Nash implementability. The next striking result is due to Matsushima (1988) and Abreu and Sen (1991).

We have to adapt the notion of ordinality appropriately, since the true set of alternatives is now the set of lotteries: this means that risk preferences are taken into account. Recall that environment E corresponds to state t, while E' does to t':

An SCR F is ordinal if, whenever $F(E) \neq F(E')$, there must exist an agent $i \in N$ and two lotteries $p, p' \in \triangle$ such that $u_i(p, t) \geq u_i(p', t)$ and $u_i(p', t') > u_i(p, t')$. That is, the concept of ordinality remains as before: for the social choice to change between two environments, a change in the relevant preferences must happen; in this case, these are von neumann-Morgenstern risk preferences over lotteries. In particular, the validity of Proposition 1 still holds: as long as we use game theory based on expected utility, ordinality remains a necessary condition for implementability in any game theoretic solution concept. The surprise is now that, with at least three agents, ordinality is also sufficient for virtual Nash implementability.

We shall also make the following very weak regularity assumption on environments:

An environment $E \in \mathcal{E}$ satisfies *no-total-indifference* (NTI) if no agent is indifferent among all alternatives in A. That is, for each agent $i \in N$ and each state $t \in \mathcal{T}$, there exist $a, a' \in A$ such that $a \succ_i^t a'$.

Proposition 8 *Consider environments satisfying NTI, and let $n \geq 3$. Any ordinal SCR F is virtually Nash implementable.*

Proof: Consider first SCRs F whose range is contained in $\triangle_+$. We claim that any such SCR is monotonic. Consider two states t and t'. Let $p \in F(t)$ and $p \notin F(t')$. Since F is ordinal, there exists $i \in N$ and lotteries p' and

p'' such that $u_i(p', t) \geq u_i(p'', t)$ and $u_i(p'', t') > u_i(p', t')$. Since agent i's preferences are of the expected utility form, this means that

$$\sum_{j=1}^{k}(p_j'' - p_j')u_i(a_j, t) \leq 0 \quad \text{while} \quad \sum_{j=1}^{k}(p_j'' - p_j')u_i(a_j, t') > 0.$$

Therefore, choosing $\lambda > 0$ small enough, one has:

$$\sum_{j=1}^{k}[p_j + \lambda(p_j'' - p_j')]u_i(a_j, t) \leq u_i(p, t) \quad \text{while} \quad \sum_{j=1}^{k}[p + \lambda(p_j'' - p_j')]u_i(a_j, t') > u_i(p, t'),$$

which implies that F is monotonic (notice how we used that $p \in \triangle_+$ in the last step).

By our NTI assumption, it is also clear that any F whose range is contained in $\triangle_+$ satisfies no-veto (because no lottery in $\triangle_+$ is top ranked for any agent). Therefore, by Proposition 7, F is exactly Nash implementable.

Finally, consider an SCR F whose range is not contained in $\triangle_+$. Notice that for any such F and for every $\epsilon > 0$, there exists an SCR F_ϵ with range contained in $\triangle_+$ such that $d(F(t), F_\epsilon(t)) < \epsilon$. That is, there exists an arbitrarily close Nash implementable SCR, which means that F is virtually Nash implementable.

∎

The result for the case of two agents is also quite permissive, but the existence of a "bad alternative" for both agents is necessary. This bad alternative will be used to punish deviations from non-unanimous reports of the state. The existence of such an alternative is guaranteed in economic environments, for example. To get a better feel for the notion of virtual implementation, we revisit once again the problem facing King Solomon.

Example 9 *Recall our Examples 4 and 4, where we had concluded that the Solomonic rule is not exactly Nash implementable. We will now show that it is virtually Nash implementable. To do this, we construct an explicit mechanism that King Solomon could use.*

Recall Ann (A) and Beth's (B) preferences, which are represented by utility functions in each state:

$$u_A(a, t_\alpha) > u_A(b, t_\alpha) > u_A(c, t_\alpha) > u_A(d, t_\alpha)$$
$$u_B(b, t_\alpha) > u_B(c, t_\alpha) > u_B(a, t_\alpha) > u_B(d, t_\alpha)$$

in state t_α, and

$$u_A(a, t_\beta) > u_A(c, t_\beta) > u_A(b, t_\beta) > u_A(d, t_\beta)$$
$$u_B(b, t_\beta) > u_B(a, t_\beta) > u_B(c, t_\beta) > u_B(d, t_\beta)$$

in state t_β. The specific utility values will not matter.

Consider the following mechanism $\Gamma = ((M_i)_{i=A,B}, g)$: for $i = A, B$, let $M_i = \mathcal{T} \times \mathcal{T} \times \mathbb{Z}_+$, that is, each woman is asked to report the state twice and a non-negative integer. Let a typical message sent by woman i be $m_i = (m_i^1, m_i^2, m_i^3)$. For a fixed $\epsilon \in (0, 1)$, the outcome function g is described in the next rules:

- (i) If $m_A^1 \neq m_B^1$, $g(m) = d$.

- (ii.a) If $m_A^1 = m_B^1 = m_A^2 = m_B^2 = t_\alpha$,
$$g(m) = (1 - \epsilon)a + \epsilon c.$$

- (ii.b) If $m_A^1 = m_B^1 = m_A^2 = m_B^2 = t_\beta$,
$$g(m) = (1 - \epsilon)b + \epsilon c.$$

- (iii.a) If $m_A^1 = m_B^1 = m_B^2 = t_\alpha \neq m_A^2$,
$$g(m) = (1 - \epsilon)a + \epsilon d.$$

- (iii.b) If $m_A^1 = m_B^1 = m_A^2 = t_\beta \neq m_B^2$,
$$g(m) = (1 - \epsilon)b + \epsilon d.$$

- (iv.a) If $m_A^1 = m_B^1 = m_A^2 = t_\alpha \neq m_B^2$,
$$g(m) = (1 - \epsilon)a + \epsilon[(1/2)a + (1/2)c].$$

- (iv.b) If $m_A^1 = m_B^1 = m_B^2 = t_\beta \neq m_A^2$,
$$g(m) = (1 - \epsilon)b + \epsilon[(1/2)b + (1/2)c].$$

- (v.a) If $m_A^1 = m_B^1 = t_\alpha \neq m_A^2 = m_B^2$,
$$\begin{array}{ll} g(m) = (1 - \epsilon)a + \epsilon[(3/4)a + (1/4)c] & \text{if} \quad m_A^3 \geq m_B^3, \\ g(m) = (1 - \epsilon)a + \epsilon[(3/4)c + (1/4)a] & \text{otherwise.} \end{array}$$

- (v.b) If $m_A^1 = m_B^1 = t_\beta \neq m_A^2 = m_B^2$,
$$\begin{array}{ll} g(m) = (1 - \epsilon)b + \epsilon[(3/4)b + (1/4)c] & \text{if} \quad m_B^3 \geq m_A^3, \\ g(m) = (1 - \epsilon)b + \epsilon[(3/4)c + (1/4)b] & \text{otherwise.} \end{array}$$

The reader can check that the only Nash equilibria of this mechanism occur under Rule (ii.a) in state t_α and under Rule (ii.b) in state t_β. Moreover, this is true for any $\epsilon \in (0, 1)$. Therefore, the proposed mechanism virtually Nash implements the Solomonic SCF.

5 Implementation in Bayesian Nash Equilibrium: Incomplete Information

We shall now consider environments in which the state $t = (t_1, t_2, \ldots, t_n)$ is not common knowledge among the n agents. We shall denote by T the set of states compatible with an environment, i.e., a set of states that is common knowledge among the agents. As such, $T \subseteq \mathcal{T}$, the domain of allowable states. Note that, under complete information, $T = \{t\}$, i.e., the set of states compatible with an environment is the singleton containing the state.

Let $T = \prod_{i \in N} T_i$, where T_i denotes the (finite) set of agent i's *types*. The interpretation is now that $t_i \in T_i$ describes the private information possessed by agent i. This private information will concern different aspects: (i) it may be about his own preferences, as in private values models (an art auction, where one's true valuation for the painting is one's private information), (ii) it may concern someone else's preferences, as in a common value problem (one may hold valuable information that is key to ascertain the true value of the object being transacted), or (iii) it may be about aspects other than preferences (even if there is no uncertainty about preferences, one agent may hold more information than the others regarding the future distribution of individual wealth). We will use the notation t_{-i} to denote $(t_j)_{j \neq i}$. Similarly, $T_{-i} = \prod_{j \neq i} T_j$.

Each agent has a *prior belief* –probability distribution– q_i defined on T, which may or may not be commonly shared. We make an assumption of non-redundant types: for every $i \in N$ and $t_i \in T_i$, there exists $t_{-i} \in T_{-i}$ such that $q_i(t) > 0$. For each $i \in N$ and $t_i \in T_i$, the conditional probability of $t_{-i} \in T_{-i}$, given t_i is the *posterior belief* of type t_i and it is denoted $q_i(t_{-i} \mid t_i)$. Let $T^* \subseteq T$ be the set of states with positive probability. We assume that agents agree on the states in T^*, i.e., for all $i \in N$, $q_i(t) = 0$ if and only if $t \notin T^*$.

The pair of objects consisting of sets of types and prior beliefs is referred to as an *information structure*. As defined, we are allowing a great deal of asymmetric information held by the agents in our model (note that the complete information structure consists of a singleton set of states and a degenerate distribution for each agent that assigns probability 1 to that state).

Recall that A denotes the (finite) set of social alternatives or outcomes, and $\triangle$ the set of lotteries over A.

Given agent i's state t utility function $u_i(\cdot, t) : \triangle \times T \mapsto \mathbb{R}$,[31] the *(interim/conditional) expected utility* of agent i of type t_i corresponding to an SCF $f : T \mapsto \triangle$ is defined as:

$$U_i(f|t_i) \equiv \sum_{t'_{-i} \in T_{-i}} q_i(t'_{-i}|t_i) u_i(f(t'_{-i}, t_i), (t'_{-i}, t_i)).$$

An *environment with incomplete information* is a list $E = \langle N, A, (u_i, T_i, q_i)_{i \in N} \rangle$. A domain is a class of environments that fix the set A of social alternatives, where each environment is common knowledge among the agents.[32]

The only informational difference between the agents and the planner is that each agent i has received his private information t_i (has "learned his type"). In contrast, while the planner knows that if agent i has received the private information t_i, agent i's posterior belief is $q_i(t_{-i} \mid t_i)$ and his conditional expected utility is $U_i(\cdot \mid t_i)$, she does not know the true t_i observed by agent i. Since the social choice that the planner wishes to make in general will depend on the agents' private information, the problem is once again when and how the planner will be able to elicit it.

For simplicity in the presentation, we shall consider only single-valued rules. A *social choice function* (SCF) f is a mapping $f : T \mapsto A$.[33] Let $\mathcal{F}$ denote the set of SCFs. Again, random SCFs map into $\triangle$, and we will use them later, when we talk about virtual implementation.

Two SCFs f and h are *equivalent* ($f \approx h$) if $f(t) = h(t)$ for every $t \in T^*$ (see Jackson (1991) for a discussion on equivalent rules).

We shall work with economic environments, as already defined in Section 4. Actually we could weaken that definition of an economic environment and all our results would go through: we could define it to be one in which there exist at least two agents with different top-ranked alternatives and there is no-total-indifference.[34]

Consider a *mechanism* $\Gamma = ((M_i)_{i \in N}, g)$ imposed on an incomplete information environment E. Note that in the present environments, a mechanism induces a game of incomplete information. Recall that a *direct mechanism*

---

[31]As the reader will have noticed, the domain of the utility function has already been extended to take account of lotteries. On the other hand, making T instead of $\mathcal{T}$ the second part of the domain is without loss of generality, as will become clear in the sequel.

[32]Just like under complete information, it will not be necessary to endow the domain of environments with a Bayesian structure, i.e., it is not important what priors on the different environments are held by the designer.

[33]Given that we assume that T itself is common knowledge among the agents, there is no loss of generality in being concerned with SCFs whose domain is T instead of $\mathcal{T}$.

[34]This weaker definition would cover the example of King Solomon.

for an SCF f is one with $M_i = T_i$ for all $i \in N$ and whose outcome function is f itself.

A *Bayesian Nash equilibrium* of Γ is a profile of strategies $\sigma^* = (\sigma_i^*)_{i \in N}$ where $\sigma_i^* : T_i \mapsto M_i$ such that $\forall i \in N$, $\forall t_i \in T_i$,

$$U_i(g(\sigma^*)|t_i) \geq U_i(g(\sigma_{-i}^*, \sigma_i')|t_i) \quad \forall \sigma_i' : T_i \mapsto M_i.$$

Observe that a Bayesian Nash equilibrium, or simply a Bayesian equilibrium, is nothing but a Nash equilibrium of the expanded game "played by the types" of the agents.[35] Therefore, what is crucial to the concept continues to be the dual property of: (a) best responses to expectations, and (b) expectations being justified by every type's optimal play.

Denote by $\mathcal{B}(\Gamma)$ the set of Bayesian equilibria of the mechanism Γ. Let $g(\mathcal{B}(\Gamma))$ be the corresponding set of equilibrium outcomes.

An SCF f is *Bayesian implementable* if there exists a mechanism $\Gamma = ((M_i)_{i \in N}, g)$ such that $g(\mathcal{B}(\Gamma)) \approx f$.

Note that we will continue to require full implementability. Historically, though, the literature began by requiring only truthful implementability, as we define next. This led to the fundamental notion of incentive compatibility, found in early works like Dasgupta, Hammond and Maskin (1979), Myerson (1979, 1981), D'Aspremont and Gerard-Varet (1979), and Harris and Townsend (1981).

We shall say that an SCF f is *truthfully implementable* or *incentive compatible* if truth-telling is a Bayesian equilibrium of the direct mechanism associated with f, i.e., if for every $i \in N$ and for every $t_i \in T_i$,

$$\sum_{t_{-i} \in T_{-i}} q_i(t_{-i}|t_i) u_i(f(t_i, t_{-i}), (t_i, t_{-i})) \geq \sum_{t_{-i} \in T_{-i}} q_i(t_{-i}|t_i) u_i(f(t_i', t_{-i}), (t_i, t_{-i}))$$
$\forall t_i' \in T_i.$

That is, an SCF f is incentive compatible whenever, if one expects the other agents to be truthful, one does not have an incentive to misrepresent his private information and report t_i' to the designer when one's true type is t_i. A fundamental result in the theory is known as the *revelation principle* (e.g., Myerson (1991)), and it justifies the central position occupied in the theory of incentives by direct mechanisms. The reason is that any Bayesian equilibrium of any game of incomplete information is outcome-equivalent to

---

[35] In fact, there are game theorists that object to this change of name. They would argue that one should still refer to this concept as Nash equilibrium. I sympathize with this view, but I will follow the majority and will retain the adjective "Bayesian" in it, to emphasize the incomplete information component. Hopefully, Reverend Bayes will not be too annoyed at the game theory community because of this.

the truth-telling equilibrium of a suitably defined direct mechanism. Therefore, by exploring the class of direct mechanisms, one is able to track down all possible equilibrium behavior that could take place in any game imposed over a given set of outcomes. For our purposes, the relevance of incentive compatibility is given by the following result.

Proposition 9 *If f is a Bayesian implementable SCF, there exists an SCF $\hat{f}$, $\hat{f} \approx f$, that is incentive compatible.*

Proof: Let f be Bayesian implementable. Therefore, there exists a mechanism $\Gamma = ((M_i)_{i \in N}, g)$ and a Bayesian equilibrium σ of Γ such that $g(\sigma(t)) = f(t)$ for every $t \in T^*$. By the revelation principle, since σ is a Bayesian equilibrium, one has that for all $i \in N$ and for all $t_i \in T_i$,

$$U_i(g(\sigma)|t_i) \geq U_i(g(m_i', \sigma_{-i}(t_{-i}))|t_i) \quad \forall m_i' \in M_i.$$

This means that the SCF $g(\sigma)$ is incentive compatible, but $g(\sigma) \approx f$.

∎

Thus, incentive compatibility is a real constraint to the set of Bayesian implementable rules.[36] The intuition is straightforward: an SCF that is not incentive compatible is never to be realized, because even if the other agents are expected to behave truthfully, there are incentives for at least one agent to behave as if his private information were diffferent.

We now investigate whether there are other necessary conditions for Bayesian implementability. The answer will turn out to be "yes," and one way to understand why is to observe that direct mechanisms associated with incentive compatible SCFs typically have the problem of multiple equilibria.[37] These considerations were first made in Postlewaite and Schmeidler (1986), and further developed in Palfrey and Srivastava (1989a), Mookherjee and Reichelstein (1990) and Jackson (1991), leading to the identification of a new condition. Before we present it, we need to go over some definitions.

---

[36]See Example 7 and, to make it an environment with incomplete information properly speaking, suppose agent j's preferences were his private information. Then, the SCF that assigns the Walrasian allocation in each state is not incentive compatible. Strenghthening this point, important impossibility results associated with incentive compatibility can be obtained (e.g., Myerson and Satterthwaite (1983)).

[37]This may happen already under complete information: suppose there are two agents and two states in T^*, $t = (t_1, t_2)$ and $t' = (t_1', t_2')$. Let $A = \{a, b, c\}$, and for $i = 1, 2$ $u_i(c, t) = u_i(c, t') = 0$, $u_i(a, t) = u_i(b, t') = 1$, $u_i(b, t) = u_i(a, t') = 2$. Let $f(t) = a$, $f(t') = b$ and $f(t_1, t_2') = f(t_1', t_2) = c$. Note that f is incentive compatible. However, the lying profile is also an equilibrium of the direct mechanism, and moreover, it dominates the truth-telling one. With examples like this, one is led to consider the multiple equilibrium problem seriously.

Consider a strategy in a direct mechanism for agent i, i.e., a mapping $\alpha_i = (\alpha_i(t_i))_{t_i \in T_i} : T_i \mapsto T_i$. A *deception* $\alpha = (\alpha_i)_{i \in N}$ is a collection of such mappings where at least one differs from the identity mapping. That is, when agents are using a deception, at least one type of one agent is lying to the planner.

Given an SCF f and a deception α, let $[f \circ \alpha]$ denote the following SCF: $[f \circ \alpha](t) = f(\alpha(t))$ for every $t \in T$. That is, $[f \circ \alpha]$ is the SCF that would be implemented if the planner wanted to implement f but the agents were to use the deception α: then, in each state t, instead of realizing $f(t)$, the outcome $f(\alpha(t))$ would result.

Finally, for a type $t_i \in T_i$, an SCF f and a deception α, let $f_{\alpha_i(t_i)}(t') = f(t'_{-i}, \alpha_i(t_i))$ for all $t' \in T$. That is, the SCF $f_{\alpha_i(t_i)}$ is what would be implemented if the planner wished to implement f, all agents other than i were to be truthful, and agent i would report that his type is $\alpha_i(t_i)$.

We shall say that an SCF f is *Bayesian monotonic* if for any deception α, whenever $f \circ \alpha \not\approx f$, there exist $i \in N$, $t_i \in T_i$ and an SCF y such that

$$U_i(y \circ \alpha \mid t_i) > U_i(f \circ \alpha \mid t_i) \text{ while } U_i(f \mid t'_i) \geq U_i(y_{\alpha_i(t_i)} \mid t'_i), \forall t'_i \in T_i. \quad (*)$$

In the spirit of monotonicity, Bayesian monotonicity justifies a change in the social choice on the basis of a preference reversal around it. Specifically, if f is the social choice, but α is a deception that undermines it (in the sense that $f \circ \alpha$ is no longer socially optimal), there must exist an SCF y and a type t_i that prefers y to f if the deception α is used, while at the same time y is not better than f for any type in T_i when every agent is truthful. The import of the condition is given by the following result.

Proposition 10 *If f is a Bayesian implementable SCF, there exists an SCF $\hat{f}$, $\hat{f} \approx f$, that is Bayesian monotonic.*

Proof: Without loss of generality, let $\hat{f} = f$. Since f is Bayesian implementable, there exists a mechanism $\Gamma = ((M_i)_{i \in N}, g)$ and a Bayesian equilibrium σ of Γ such that $g(\sigma) = f$.

Since σ is a Bayesian equilibrium, it follows that for every $i \in N$ and for every $t_i \in T_i$, the set of outcomes $\{g(m'_i, \sigma_{-i}(t_{-i}))\}_{m'_i \in M_i}$ must be contained in the set Φ_i consisting of all the SCFs y satisfying that $U_i(f|t_i) \geq U_i(y_{\beta_i(t_i)}|t_i)$ for every $(\beta_i(t_i))_{t_i \in T_i} : T_i \mapsto T_i$ for all $t_i \in T_i$. That is, no unilateral deviation on the part of type t_i, including those in which he would pretend to be type $\beta_i(t_i)$, is profitable from his point of view.

Now fix a deception α that undermines f, i.e., $f \circ \alpha \not\approx f$. Since f is Bayesian implementable, the strategy profile $\sigma \circ \alpha$, where for every $i \in N$

each type t_i behaves as in $\sigma_i(\alpha_i(t_i))$ cannot be a Bayesian equilibrium. That is, there must exist a type t_i who can force an outcome $y \circ \alpha$ via a unilateral deviation from $\sigma \circ \alpha$ and that he prefers to $f \circ \alpha$. Furthermore, if $y \circ \alpha$ can be so induced, then the corresponding y could also be induced via a unilateral deviation from σ, and thus it must belong to Φ_i. But this is the preference reversal asked in equation (*) by Bayesian monotonicity. ∎

The following characterization of Bayesian implementable SCFs for exchange economies (recall Example 3) is related to a result due to Jackson (1991).[38] We present a simple direct proof of this result.

Proposition 11 *Consider an exchange economy and an SCF f that never assigns the zero bundle to any agent. If there exists $\hat{f} \approx f$ that is incentive compatible and Bayesian monotonic, then f is Bayesian implementable.*

[39]

Proof: Without loss of generality, let $\hat{f} = f$. We shall assume that the aggregate endowment is the only top ranked bundle of goods for each agent in each state. The proof constructs a canonical mechanism that implements f in Bayesian equilibrium, whenever f satisfies the required properties. We describe the mechanism presently; note that it is augmented with respect to the direct mechanism for f.

Consider the mechanism $\Gamma = ((M_i)_{i \in N}, g)$, where $M_i = T_i \times \mathcal{F} \times \mathbb{Z}_+$, i.e., each agent is asked to report his type, an SCF and a non-negative integer. The outcome function g is as follows:

- (i) If for all $i \in N$, $m_i = (t_i, f, 0)$, $g(m) = f(t)$, where $t = (t_1, t_2, \dots, t_n)$.

- (ii) If for all $j \neq i$, $m_j = (t_j, f, 0)$ and $m_i = (t_i', y, z_i)$, we can have two cases:

  - (ii.a) If for all t_i and for all $\alpha_i(t_i)$, $U_i(y_{\alpha_i(t_i)}|t_i) \leq U_i(f|t_i)$, $g(m) = y(t_i', t_{-i})$.
  - (ii.b) Otherwise, $g(m) = f(t_i', t_{-i})$.

---

[38] Jackson (1991) provides a characterization of set-valued rules when there are at least three agents. For those rules, in addition to incentive compatibility and Bayesian monotonicity, there is a third condition that is also necessary and sufficient: the set-valued rule must be closed under the concatenation of different common knowledge events. This issue does not arise if one considers only SCFs.

[39] The zero bundle condition on f is added to give this simple proof. In general, such a requirement is not part of the characterization.

- (iii) In all other cases, the total endowment of the economy is awarded to the agent of smallest index among those who announce the largest integer.

We now prove the proposition. First, we show that f can be supported by a Bayesian equilibrium of this mechanism. Consider the following strategy profile, where agent i of type t_i announces $m_i(t_i) = (t_i, f, 0)$. If this profile is played, Rule (i) is imposed and the outcome $f(t)$ results when the true state is t.

By incentive compatibility, agent i of type t_i would not want to change his type report (in this case, the outcome would still fall under Rule (i). Changing his integer would produce an outcome under Rule (ii) and the equilibrium outcome $f(t)$ would result. Finally, he could change his announced SCF to y, but then y would be implemented only when for all types of player i, y is not preferred to f. The same goes if he were to deviate modifying his announced type and SCF: some $y_{\alpha_i(t_i)}$ would result only when it is not better than f for any type of agent i. Therefore, the proposed profile is a Bayesian equilibrium, and its outcome is f.

In equilibrium, the outcome will never fall under Rule (iii): a type who is not winning the integer game has an incentive to announce an integer larger than those announced by every type of every agent. This will increase his expected utility (note that f never assigns the aggregate endowment to any agent).

Rule (ii) also conflicts with equilibrium: one of the agents $j \neq i$ can deviate and announce an SCF with all resources awarded to him, as well as the highest integer across all types. This would increase his expected utility.

Thus, all equilibrium strategies fall under Rule (i), i.e., f is unanimously announced and all agents announce the integer 0. However, we could have that type t_i of agent i announces type $\alpha_i(t_i)$ instead of a truth telling report. When the true type profile is t, this would result in the outcome $f \circ \alpha$ being implemented.

If $f \circ \alpha \approx f$, we are done, i.e., we have multiple equilibria, but with the same outcome. Thus, suppose that $f \circ \alpha \not\approx f$. Because f is Bayesian monotonic, there exists a type t_i and an SCF y exhibiting the preference reversal in equation (*). Then, let type t_i deviate from his equilibrium message, and instead send $m_i'(t_i) = (\alpha_i(t_i), y, 0)$ (his alledged equilibrium message was $(\alpha_i(t_i), f, 0)$). The result is that $y \circ \alpha$ is implemented, instead of $f \circ \alpha$; to see this, note that Rule (ii.a) is being used thanks to the properties of y. However, this is a profitable deviation for type t_i, which contradicts that this profile is a Bayesian equilibrium.

■

Therefore, full Bayesian implementability is equivalent to incentive compatibility and Bayesian monotonicity in economic environments. We already saw that incentive compatibility is sometimes a real constraint that may crowd out interesting social rules. Our next attempt is to get a better understanding of the strength of Bayesian monotonicity (see Palfrey and Srivastava (1987), in which interesting rules are shown to satisfy Bayesian monotonicity in certain domains).[40] In general, though, Bayesian monotonicity turns out to be a very demanding condition, representing a serious obstacle to implementability. This is argued in the following example.

Example 10 [41] *Consider an exchange economy as those described in Examples 3 and 7. Let $N = \{1, 2, 3, 4\}$. There is a single commodity -money- and all consumers have 1 dollar as endowment in each state. The set of allocations may again be infinite or finite: either we admit infinitely divisible money, or only all feasible assignments of dollars and cents to the four consumers. The sets of types are $T_k = \{t_k, t'_k, t''_k\}$ for $k = 1, 2$, while $T_j = \{t_j, t'_j\}$ for $j = 3, 4$. There are only three states which arise with positive probability: $T^* = \{t, t', t''\}$, where $t = (t_1, t_2, t_3, t_4)$, $t' = (t'_1, t'_2, t'_3, t'_4)$ and $t'' = (t''_1, t''_2, t'_3, t'_4)$. Agents 1 and 2 are fully informed, so that for $k = 1, 2$, the posterior probability distributions are:*

$$q_k(t|t_k) = q_k(t'|t'_k) = q_k(t''|t''_k) = 1, \qquad k = 1, 2.$$

Agents 3 and 4 are fully informed only when they are of type t_j, i.e., for $j = 3, 4$, $q_j(t|t_j) = 1$, but

$$q_3(t'|t'_3) = 0.25, \qquad q_3(t''|t'_3) = 0.75,$$

$$q_4(t'|t'_4) = 0.75, \qquad q_4(t''|t'_4) = 0.25.$$

The utility functions are as follows:

$$u_i(x, s) = x_i^{\lambda_i(s)}, \quad \lambda_i(s) \in (0, 1) \quad \forall s \in T, \quad \forall i \in N.$$

Note first that incentive compatibility is not a constraint in this environment. The presence of at least two fully informed agents in each state guarantees that truth-telling is always a Bayesian equilibrium of the direct

---

[40]See also Matsushima (1993), where it is shown that Bayesian monotonicity is a trivial condition in environments with quasilinear utilities.

[41]Variants of this example were developed in Palfrey and Srivastava (1987), Chakravorti (1992) and Serrano and Vohra (2001).

mechanism for any SCF: one can punish unilateral deviations from truth-telling with the zero bundle for every agent.

In contrast, Bayesian monotonicity is very restrictive in the present example. Let f be an SCF such that for some $s \in T^*$, $s \neq t$, $f(s) \neq f(t)$. Then f is not Bayesian monotonic. To see this, consider a deception α such that every type of every agent reports that he is of type t_i: $\alpha_i(s_i) = t_i$ for every $s_i \in T_i$ for all $i \in N$. For this deception, $f \circ \alpha \not\succapprox f$ since $f \circ \alpha$ is a constant SCF that assigns $f(t)$ in every state. Bayesian monotonicity requires then the existence of a type s_i and an SCF y exhibiting the preference reversal of equation (*). Since $f \circ \alpha$ specifies $f(t)$ in every state, it follows that

$$\text{for all } i \text{ and all } s \in T, \ U_i(f \circ \alpha \mid s_i) = \sum_{s'_{-i} \in T_{-i}} q_i(s'_{-i}|s_i) u_i((f(t), (s'_{-i}, s_i))).$$

Since for each i, $u_i(\cdot, s)$ represents the same ordinal preferences in each state $s \in T$, it follows that for all i and $s \in T$, whenever $U_i(f \mid \alpha_i(s_i)) \geq U_i(y \mid \alpha_i(s_i))$, $U_i(f \circ \alpha \mid s_i) \geq U_i(y \circ \alpha \mid s_i)$. This is sufficient to assert that one cannot find a reversal as specified in (*). Thus, Bayesian monotonicity implies that f must be constant over T^*.

To finish the example, we remark that, if one wants to take into account the property rights implied by the initial endowments, one should consider SCFs in which each type ends up with an interim utility no lower than 1. In addition, if one adds efficiency considerations to the social optimum, such an SCF cannot be constant: while it must prescribe the endowment point in state t, for insurance reasons, types t'_3 and t'_4 should write contracts between them to trade part of their endowments. It follows that no such SCF is Bayesian implementable. That is, any mechanism that has a Bayesian equilibrium supporting the desired rule will also have an equilibrium in which every type pretends to be of type t_i.

Therefore, in environments like those in Example 10, exact Bayesian implementability reduces to constant SCFs (and of course, a planner does not need any sophisticated theory of implementation to impose those). We remark also that Bayesian monotonicity is always a necessary condition for Bayesian implementability, even in two-agent environments. The sufficiency result showed in Proposition 11 also holds for two-agent environments: this is done because the mechanism relies on f, the SCF to be implemented, being announced in the equilibrium strategy profile. Thus, the difficulty of identifying a deviator from the "right" announcement does not arise, even if one considers only two-agent settings.

40

5.1 Virtual Implementation in Bayesian Equilibrium

Given the limitations of Bayesian implementability and the remarkable positive results of virtual Nash implementation, it makes sense to investigate the virtual approach to implementation in Bayesian environments. This was done in Abreu and Matsushima (1992b), Duggan (1997) and Serrano and Vohra (2001, 2002). We begin the subsection with a few more definitions.

Recall that $\triangle$ denotes the set of probability distributions over the set A of social alternatives. A (random) *social choice function* (SCF) f is a function $f : T \mapsto \triangle$. A (random) *mechanism* $\Gamma = ((M_i)_{i \in N}, g)$ describes a message set M_i for agent i and an outcome function $g : \prod_{i \in N} M_i \mapsto \triangle$. Recall that $\mathcal{B}(\Gamma)$ is the set of Bayesian equilibria of Γ and $g(\mathcal{B}(\Gamma))$ the corresponding set of outcomes.

Consider the following metric on SCFs:

$$d(f, h) = \max\{|f(a \mid t) - h(a \mid t)| \mid t \in T^*, a \in A\}.$$

An SCF f is *virtually Bayesian implementable* if $\forall \epsilon > 0$ there exists an SCF f^ϵ such that $d(f, f^\epsilon) < \epsilon$ and f^ϵ is exactly Bayesian implementable.

In investigating the conditions that will characterize virtual Bayesian implementability, note first that a new application of the revelation principle implies that incentive compatibility continues to be a necessary condition (the proof is very similar to that of Proposition 9).[42] In addition, the following condition has been shown to be also necessary, as well as sufficient in conjunction with incentive compatibility (Serrano and Vohra (2002)).

An SCF f is *virtually monotonic* if for every deception α, whenever $f \not\approx f \circ \alpha$, there exists $i \in N$, $t_i \in T_i$, an incentive compatible SCF x and an SCF y such that

$$U_i(y \circ \alpha \mid t_i) > U_i(x \circ \alpha \mid t_i) \text{ while } U_i(x \mid t_i') \geq U_i(y_{\alpha_i(t_i)} \mid t_i'), \forall t_i' \in T_i. \quad (**)$$

It is instructive to compare virtual monotonicity with Bayesian monotonicity. The difference between the two is that the preference reversal in $(**)$ can happen around any incentive compatible SCF x in the environment, not necessarily around f as was required in equation $(*)$. Clearly, in the class of incentive compatible SCFs, virtual monotonicity is weaker than Bayesian monotonicity.

---

[42]Abreu and Matsushima (1992b) and Duggan (1997) provide two independent conditions that, together with incentive compatibility, are sufficient (these are measurability and incentive consistency, respectively). These two conditions are permissive, but they are not necessary; in fact, they are sometimes stronger than Bayesian monotonicity (Serrano and Vohra (2001)).

Instead of stating and proving the characterization theorem, in order to convey the very positive nature of the results achieved with virtual Bayesian implementation, we introduce the following condition on an environment.

An environment E satisfies *type diversity* if there do not exist $i \in N$, $t_i, t'_i \in T_i$, $t_i \neq t'_i$, $\beta \in \mathbb{R}_{++}$ and $\gamma \in \mathbb{R}$ such that

$$U_i(a|t_i) = \beta U_i(a|t'_i) + \gamma \quad \forall a \in A.$$

This condition has a simple interpretation: it requires that the interim preferences over pure alternatives of different types of an agent be different. In the space of preferences, type diversity is generically satisfied by an environment if there are at least three alternatives.[43]

Proposition 12 *In economic environments satisfying type diversity, an SCF f is virtually Bayesian implementable if and only if there exists an equivalent SCF $\hat{f}$ that is incentive compatible.*

Proof: Without loss of generality, let $\hat{f} = f$. We already know that incentive compatibility is always a necessary condition for virtual Bayesian implementability. It remains to show sufficiency.

First of all, it can be shown that type diversity guarantees the existence of random SCFs $l_i(t_i)$ that are constant over T and that satisfy that

$$U_i(l_i(t_i)|t_i) > U_i(l_i(t'_i)|t_i) \quad \forall t'_i \in T_i, \forall t_i \in T_i, t'_i \neq t_i.$$

In addition, our assumption of economic environments ensures no-total-indifference.

We now construct a mechanism that virtually implements f in Bayesian equilibrium. Let $\Gamma = ((M_i)_{i \in N}, g)$, where $M_i = T_i \times T_i \times A \times \mathbb{Z}_+$. A typical message sent by agent i will be denoted $m_i = (t_i^1, t_i^2, a_i, z_i)$. A strategy of agent i is denoted σ_i, where $\sigma_i : T_i \mapsto M_i$. Let $t^1 = (t_1^1, t_2^1, \ldots, t_n^1)$ be the profile of first type reports, and $t^2 = (t_1^2, t_2^2, \ldots, t_n^2)$ be the profile of types reported in second place by each agent.

Let $L = \{((l_i(t_i))_{t_i \in T_i})_{i \in N})\}$ be the collection of all (constant) SCFs implied by type diversity. Let $L(t) = \{(l_i(t_i))_{i \in N}\}$ and $\bar{l} = \frac{1}{|L|} \sum_{l \in L} l$. Note that the SCF $\bar{l}$ is constant over all $t \in T$. For any $a \in A$ and $\lambda \in [0,1]$ let $a(\lambda) = \lambda a + (1 - \lambda)\bar{l}$.

For $\epsilon \in (0,1)$, the outcome function (ϵ-close to f) is defined as follows.

---

[43]In environments satisfying type diversity, Abreu-Matsushima's measurability, Duggan's incentive consistency and Serrano-Vohra's virtual monotonicity are satisfied by every SCF. Therefore, the proof below does not need to rely on any of these conditions. We borrow this proof from an earlier draft of Serrano and Vohra (2002).

- (i) If there exists $j \in N$ such that for all $i \neq j$, $m_i = (t_i, t_i, a_i, 0)$, and $z_j = 0$:
$$g(m) = (1 - \epsilon)f(t^1) + \frac{\epsilon}{2n}\sum_{i \in N} l_i(t_i^1) + \frac{\epsilon}{2}\bar{l}.$$

- (ii) If there exists $j \in N$ such that for all $i \neq j$, $m_i = (t_i, t_i, a_i, 0)$, and $z_j > 0$:
$$g(m) = (1 - \epsilon)f(t^1) + \frac{\epsilon}{2n}\sum_{i \in N} l_i(t_i^2) + \frac{\epsilon}{2}\bar{l}.$$

- (iii) Otherwise, denoting by h the agent with the lowest index among those who announce the highest integer:
$$g(m) = (1 - \epsilon)f(t^1) + \frac{\epsilon}{2n}\sum_{i \in N} l_i(t_i^2) + \frac{\epsilon}{2}a_h\left(\frac{z_h}{z_h + 1}\right).$$

A strategy profile where for each $i \in N$ and each $t_i \in T_i$, $m_i(t_i) = (t_i, t_i, a_i, 0)$ is a Bayesian equilibrium of the mechanism Γ. To see this, note that changing the first type report is not a profitable deviation because f is incentive compatible and because $U_i(l_i(t_i) \mid t_i) > U_i(l_i(t_i') \mid t_i)$ for any $t_i' \neq t_i$. The latter condition also implies that it is not profitable to change the announcement of the number and the second type report. Changing the announced alternative a_i does not alter the outcome. Therefore, the proposed strategy profile is a Bayesian equilibrium of Γ.

Next, note that using the standard argument for the integer game, an equilibrium cannot happen under Rule (iii), or under Rules (II) or (i) when an agent is announcing different types. In the last two cases, the reason is simply that the integer game can be triggered by a unilateral deviation.

Finally, there cannot be an equilibrium of Γ under Rule (i) where all agents report the same type twice and the number 0, but where these type reports are not truthful. Otherwise, given the properties of the SCFs $l_i(t_i)$, any agent who is not reporting his true type could increase his expected utility by reporting his true type in second place and choosing a number greater than 0. Thus every equilibrium corresponds to case (i), for all $t \in T^*$, with all agents reporting their types truthfully.

∎

Example 11 *Consider the exchange economy of Example **??**, and note that type diversity is satisfied whenever the constants $\lambda_i(s)$ are all different. Thus, for almost every environment in this example, every SCF is virtually Bayesian implementable.*

6 Concluding Remarks and Other Topics

The results reported in this survey identify the class of social choice rules that can be implementable under complete and incomplete information when the agents are assumed to play either dominant or equilibrium strategies. We have learned that implementation in dominant strategies is very restrictive if the planner has no information about agents' preferences, so that she must consider the unrestricted domain of preferences. On the other hand, positive results emerge on restricted domains, leading to the identification of interesting SCRs, whose implementation is therefore particularly robust. For implementation in Nash equilibrium, the key condition is monotonicity, which allows for more permissive results, especially when set-valued goals are considered. If one turns to incomplete information, implementation in Bayesian equilibrium is more restrictive: apart from incentive compatibility, there are obstacles related to Bayesian monotonicity, which may be quite demanding.

The virtual approach to implementation stands out as an extremely permissive theory. Almost always one can virtually implement any incentive compatible social rule. This approach may be especially successful in applications, where one should allow some degree of freedom in making "mistakes" in the implementation of the rule and in the exact specification of agents' preferences. In dealing with specific applications, one should strive for the design of mechanisms less abstract than the ones constructed in this survey. Modern technology makes this concern less pressing, though, at least if one expects this theory to inspire the design of mechanisms to be played over a computer network: for example, random devices, modulo games and alike can easily be implemented via computer protocols.

We close with a brief discussion of other important topics that have been touched upon by the theory of implementation. The list is not meant to be comprehensive, and apoligies are due to the authors whose work I am ignoring.

Other solution concepts: One can adopt different forms of behavior of agents other than dominance and Nash equilibrium. This would correspond to employing other game theoretic solution concepts, including undominated equilibrium (Palfrey and Srivastava (1989b, 1991), Jackson, Palfrey and Srivastava (1994), Sjostrom (1994)), dominance solvability and iterative elimination of dominated strategies (Moulin (1979), Abreu and Matsushima (1994)), trembling-hand perfect Nash equilibrium (Sjostrom (1993)), strong Nash equilibrium (Maskin (1979)). The key issue analyzed here is how the necessary conditions identified for these concepts compare to monotonicity.

Although one cannot make a general statement, one can sometimes get more permissive results.

Commitment issues: The theory presented in the current paper assumes that the designer can commit to the mechanism she proposes, or if there is no designer, that the outside enforcer will be able to enforce the outcome prescribed in the mechanism, whatever this is. Suppose, however, that some of the outcomes prescribed are not efficient. Then, agents have an incentive to renegotiate them before they go to the enforcer. The constraints that renegotiation imposes on implementability were studied in Maskin and Moore (1999) (see Rubinstein and Wolinsky (1992) for a different approach). Some other times the outcome prescribed leaves agents worse than they started with, raising the issue of whether they would want to participate in such a mechanism (Jackson and Palfrey (2001) study this problem of voluntary implementation). Finally, the possibility of having the planner as a player was introduced in Baliga, Corchón and Sjostrom (1997) and Baliga and Sjostrom (1999).

Simplicity and robustness: One criticism raised against the canonical mechanism is that it may be too complex. If the designer has good information about certain aspects of preferences, (e.g., the rates at which agents want to substitute the consumption of one good for another), it is possible to write simpler mechanisms that rely on a small number of parameters, such as prices of marketted goods (e.g., Dutta, Sen and Vohra (1995)). One can argue that relying on simpler mechanisms leads to more robust implementation. Other ways to model robustness have taken the form of double implementation (i.e., the mechanism is able to implement in two –or more– solution concepts, as in Yamato (1993)), or continuous implementation (by constructing continuous outcome functions, as in Postlewaite and Wettstein (1989)).

Applications: As we said above, less abstract, more realistic mechanisms can be designed when one deals with a specific application. One important research agenda in game theory, known as the Nash program, aims to design mechanisms that implement specific cooperative solutions. In doing so, an explicit desideratum of such mechanisms is their "appeal" in terms of realistic ways for agents to interact; see the celebrated implementation of the Nash bargaining solution via Rubinstein's game of alternating offers (Nash (1950b), Rubinstein (1982), Binmore, Rubinstein and Wolinsky (1986)).[44] It is especially remarkable when one finds successful uses of mechanisms in

---

[44]Serrano (1997) proposes a general approach to understand the Nash program under the theory of implementation.

the real world. Mechanisms have sometimes been put in place by intended expert design (e.g., the design of matching markets for hospitals and residents; see Roth (2002)). Perhaps even more amazing is that mechanisms are sometimes the consequence of wise ancient tradition (Cabrales, Calvó-Armengol and Jackson (2003) report on a cooperative of farmers in Andorra; for centuries, these families of farmers have been making payments to the cooperative to provide fire insurance, and it turns out that the mechanism used Nash-implements an almost efficient allocation).

Bounded rationality: Some early attempts were made to understand the properties of learning Nash equilibria in mechanisms that implement Walrasian allocations (Jordan (1986)). More recently, Cabrales (1999), and Cabrales and Ponti (2000) analyze the convergence properties of several dynamic processes in some of the canonical mechanisms. Some of their conclusions are quite interesting. For example, it turns out that the type of learning process specified in Cabrales (1999) for the canonical mechanism of Nash implementation always converges to the Nash equilibria of the mechanism, thereby dispelling the claim that this mechanism is far too complex; see also Sandholm (2002) for an interesting application of evolutionary implementation. Finally, Eliaz (2002) studies a model in which a number of agents are faulty (their behavior is totally unpredictable); even their identity is unknown to the designer. Interestingly, his results can also be related to monotonicity.

References

Abreu, D. and H. Matsushima (1992a), "Virtual Implementation in Iteratively Undominated Strategies: Complete Information," *Econometrica* **60**, 993-1008.

Abreu, D. and H. Matsushima (1992b), "Virtual Implementation in Iteratively Undominated Strategies: Incomplete Information," mimeo, Princeton University.

Abreu, D. and H. Matsushima (1994), "Exact Implementation," *Journal of Economic Theory* **64**, 1-19.

Abreu, D. and A. Sen [1990], "Subgame Perfect Implementation: A Necessary and Almost Sufficient Condition," *Journal of Economic Theory* **50**, 285-299.

Abreu, D. and A. Sen (1991), "Virtual Implementation in Nash Equilibrium," *Econometrica* **59**, 997-1021.

Akerloff, G. A. (1970), "The Market for "Lemons": Quality Uncertainty and the Market Mechanism," *Quarterly Journal of Economics* **84**, 488-500.

Arrow, K. J. (1951), *Social Choice and Individual Values*, New York, Wiley.

Aumann, R. J. (1976), "Agreeing to Disagree," *Annals of Statistics* **4**, 1236-1239.

Baliga, S. (1999), "Implementation in Economic Environments with Incomplete Information: the Use of Multi-Stage Games," *Games and Economic Behavior*

27, 173-183.

Baliga, S. and T. Sjostrom (1999), "Interactive Implementation," *Games and Economic Behavior* **27**, 38-63.

Baliga, S., L. Corchón and T. Sjostrom (1997), "The Theory of Implementation when the Planner Is a Player," *Journal of Economic Theory* **77**, 15-33.

Barberá, S. (1977), "The Manipulation of Social Choice Mechanisms that Do not Leave too much to Chance," *Econometrica* **45**, 1573-1588.

Barberá, S. (1983), "Strategy-Proofness and Pivotal Voters: a Direct Proof of the Gibbard-Satterthwaite Theorem," *International Economic Review* **24**, 413-417.

Barberá, S. and B. Peleg (1990), "Strategy-Proof Voting Schemes with Continuous Preferences," *Social Choice and Welfare* **7**, 31-38.

Barberá, S., A. Bogomolnia, and H. van der Stel (1998), "Strategy-Proof Probabilistic Rules for Expected Utility Maximizers," *Mathematical Social Sciences* **35**, 89-103.

Barberá, S., B. Dutta, and A. Sen (2001), "Strategy-Proof Social Choice Correspondences," *Journal of Economic Theory* **101**, 374-394.

Barberá, S., H. Sonnenschein and L. Zhou, (1991), "Voting by Committees," *Econometrica* **59**, 595-609.

Benoit, J.-P. (1999), "The Gibbard-Satterthwaite Theorem: a Simple Proof," mimeo, New York University.

Benoit, J.-P. (2002), "Strategic Manipulation of Voting Games when Lotteries and Ties Are Permitted," *Journal of Economic Theory* **102**, 421-436.

Bergin, J. and A. Sen (1998), "Extensive Form Implementation in Incomplete Information Environments," *Journal of Economic Theory* **80**, 222-256.

Binmore, K. G., A. Rubinstein and A. Wolinsky (1986), "The Nash Bargaining Solution in Economic Modeling," *RAND Journal of Economics* **17**, 176-188.

Brusco, S. (1995), "Perfect Bayesian Implementation," *Economic Theory* **5**, 419-444.

Cabrales, A. (1999), "Adaptive Dynamics and the Implementation Problem with Complete Information," *Journal of Economic Theory* **86**, 159-184.

Cabrales, A. and G. Ponti (2000), "Implementation, Elimination of Weakly Dominated Strategies and Evolutionary Dynamics," *Review of Economic Dynamics* **3**, 247-282.

Cabrales, A., A. Calvó-Armengol and M. O. Jackson (2003), "La Crema: A Case Study of Mutual Fire Insurance," *Journal of Political Economy* **111**, 425-458.

Chakravorti, B. (1992), "Efficiency and Mechanisms with no Regret," *International Economic Review* **33**, 45-59.

Chen, Y. (forthcoming), "Incentive-Compatible Mechanisms for Pure Public Goods: A Survey of Experimental Literature," in C. Plott and V. Smith (eds.) *Handbook of Experimental Economics Results*, Elsevier Science.

Ching, S. and L. Zhou (2002), "Multi-Valued Strategy-Proof Social Choice Rules," *Social Choice and Welfare* **19**, 569-580.

Clarke, E. H. (1971), "Multi-Part Pricing of Public Goods," *Public Choice* **2**, 19-33.

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

Dagan, N., R. Serrano and O. Volij, (1999), "Feasible Implementation of Taxation Methods," *Review of Economic Design* **4**, 57-72.

Dasgupta, P., P. Hammond and E. Maskin (1979), "Implementation of Social Choice Rules: Some General Results on Incentive Compatibility," *Review of Economic Studies* **46**, 195-216.

D'Aspremont and Gerard-Varet (1979), "Incentives and Incomplete Information," *Journal of Public Economics* **11**, 25-45.

Duggan, J. (1997), "Virtual Bayesian Implementation," *Econometrica* **65**, 1175-1199.

Duggan, J. and T. Schwartz (2000), "Strategic Manipulability without Resoluteness or Shared Beliefs: Gibbard-Satterthwaite Generalized," *Social Choice and Welfare* **17**, 85-93.

Dutta, B. and A. Sen (1991), "Necessary and Sufficient Conditions for Two-Person Nash Implementation," *Review of Economic Studies* **58**, 121-128.

Dutta, B., A. Sen and R. Vohra (1995), "Nash Implementation through Elementary Mechanisms in Economic Environments," *Economic Design* **1**, 173-203.

Eliaz, K. (2002), "Fault-Tolerant Implementation," *Review of Economic Studies* **69**, 589-610.

Geanakoplos, J. (1996), "Three Brief Proofs of Arrow's Impossibility Theorem," mimeo, Cowles Foundation Yale University, 1996.

Gibbard, A. (1973), "Manipulation of Voting Schemes: a General Result," *Econometrica* **41**, 587-601.

Glazer, J. and A. Ma (1989), "Efficient Allocation of a Prize: King Solomon's Problem," *Games and Economic Behavior* **1**, 222-233.

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

Groves, T. (1973), "Incentives in Teams," *Econometrica* **41**, 617-631.

Harris, M. and R. Townsend (1981), "Resource Allocation with Asymmetric Information," *Econometrica* **49**, 33-64.

Hayek, F. (1945), "The Use of Knowledge in Society," *American Economic Review* **35**, 519-530.

Hong (1998), "Feasible Bayesian Implementation with State Dependent Feasible Sets," *Journal of Economic Theory*.

Hurwicz, L. (1960), "Optimality and Informational Efficiency in Resource Allocation Processes," in K. J. Arrow et al. (eds.) *Mathematical Methods in the Social Sciences*, Stanford University Press.

Hurwicz, L. (1972), "On Informationally Decentralized Systems," in *Decision and Organization*, C. B. McGuire and R. Radner (eds.), Amsterdam, North Holland.

Hurwicz, L. (1979)], "Outcome Functions Yielding Walrasian and Lindahl Allocations at Nash Equilibrium Points," *Review of Economic Studies*, **46**, 217-225.

Hurwicz, L. and D. Schmeidler (1978), "Construction of Outcome Functions Guaranteeing Existence and Pareto Optimality of Nash Equilibria," *Econometrica* **46**, 1427-1474.

Hurwicz, L., E. Maskin and A. Postlewaite (1995), "Feasible Nash Implementation of Social Choice Rules when the Designer Does not Know Endowments

or Production Sets," in Ledyard, J. O. (ed.) *The Economics of Informational Decentralization: Complexity, Efficiency and Stability*, Kluwer Academic Publishers, Amsterdam.

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

Jackson, M. O. (1992), "Implementation in Undominated Strategies: A Look at Bounded Mechanisms," *Review of Economic Studies* **59**, 757-775.

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

Jackson, M. O. and T. R. Palfrey (2001), "Voluntary Implementation," *Journal of Economic Theory* **98**, 1-25.

Jackson, M. O., T. R. Palfrey and S. Srivastava (1994), "Undominated Nash Implementation in Bounded Mechanisms," *Games and Economic Behavior* **6**, 474-501.

Jordan, J. (1986), "Instability in the Implementation of Walrasian Allocations," *Journal of Economic Theory* **39**, 301-328.

Laffont, J.-J. and E. Maskin (1982), "The Theory of Incentives: an Overview," in W. Hildebrand (ed.) *Advances in Economic Theory, IV World Congress of the Econometric Society*, Cambridge University Press.

Lange, O. (1936), "On the Theory of Socialism. Part I," *Review of Economic Studies* **3**, 53-71.

Lange, O. (1937), "On the Theory of Socialism. Part II," *Review of Economic Studies* **4**, 123-142.

Mandeville, B. (1732), *The Fable of the Bees: Private Vices, Public Benefits*, Liberty Books (1924 edition).

Mas-Colell, A. and X. Vives (1993), "Implementation in Economies with a Continuum of Agents," *Review of Economic Studies* **60**, 613-629.

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

Maskin, E. (1979), "Implementation in Strong Nash Equilibrium," in J.-J. Laffont (ed.) *Aggregation and Revelation of Preferences*, Amsterdam, North-Holland.

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 University Press.

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

Maskin, E. and J. Moore (1999), "Implementation and Renegotiation," *Review of Economic Studies* **66**, 83-114.

Maskin, E. and T. Sjostrom (2002), "Implementation Theory," in K. J. Arrow, A. Sen and K. Suzumura (eds.) *Handbook of Social Choice and Welfare (vol. I)*, New York, Elsevier Science B.V.

Matsushima, H. (1988), "A New Approach to the Implementation Problem," *Journal of Economic Theory* **45**, 128-144.

Matsushima, H. (1993), "Bayesian Monotonicity with Side Payments," *Journal of Economic Theory* **59**, 107-121.

Mirrlees, J. (1971), "An Exploration in the Theory of Optimal Income Taxation," *Review of Economic Studies* **38**, 175-208.

Mookherjee, D. and S. Reichelstein (1990), "Implementation via Augmented Revelation Mechanisms," *Review of Economic Studies* **57**, 453-475.

Moore, J. [1992], "Implementation, Contracts and Renegotiation in Environments with Complete Information," in J. J. Laffont (ed.) *Advances in Economic Theory, VI World Congress of the Econometric Society* (vol. I), Cambridge University Press, Cambridge.

Moore, J. and R. Repullo [1988], "Subgame Perfect Implementation," *Econometrica* **56**, 1191-1220.

Moore, J. and R. Repullo (1990), "Nash Implementation: A Full Characterization," *Econometrica* **58**, 1083-1100.

Moulin, H. (1979), "Dominant Solvable Voting Schemes," *Econometrica* **47**, 1337-1352.

Moulin, H. (1983), *The Strategy of Social Choice*, Amsterdam, North Holland.

Mueller, E. and M. Satterthwaite (1977), "The Equivalence of Strong Positive Association and Strategy-Proofness," *Journal of Economic Theory* **14**, 412-418.

Myerson, R. B. (1979), "Incentive Compatibility and the Bargaining Problem," *Econometrica* **47**, 61-73.

Myerson, R. B. (1981), "Optimal Auction Design," *Mathematics of Operations Research* **6**, 58-73.

Myerson, R. B. (1991), *Game Theory: Analysis of Conflict*, Harvard University Press, Cambridge.

Myerson, R. B. and M. Satterthwaite (1983), "Efficient Mechanisms for Bilateral Trading," *Journal of Economic Theory* **28**, 265-281.

Nash, J. F. (1950a), "Equilibrium Points in n-Person Games," *Proceedings of the National Academy of Sciences* **36**, 48-49.

Nash, J. F. (1950b), "The Bargaining Problem," *Econometrica* **18**, 155-162.

Osborne, M. J. and A. Rubinstein (1994), *A Course in Game Theory*, MIT Press, Cambridge.

Palfrey, T. R. (1992), "Implementation in Bayesian Equilibrium: the Multiple Equilibrium Problem in Mechanism Design," in J. J. Laffont (ed.) *Advances in Economic Theory, VI World Congress of the Econometric Society* (vol. I), Cambridge University Press, Cambridge.

Palfrey, T. R. (2002), "Implementation Theory," in R. J. Aumann and S. Hart (eds.) *Handbook of Game Theory with Economic Applications (vol. III)*, New York, Elsevier Science.

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

Palfrey, T. R. and S. Srivastava (1989a), "Implementation with Incomplete Information in Exchange Economies," *Econometrica* **57**, 115-134.

Palfrey, T. R. and S. Srivastava (1989b), "Mechanism Design with Incomplete Information: a Solution to the Implementation Problem," *Journal of Political Economy* **97**, 668-691.

Palfrey, T. R. and S. Srivastava (1991), "Nash Implementation Using Undominated Strategies," *Econometrica* **59**, 479-501.

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

Postlewaite, A. and D. Wettstein (1989), "Feasible and Continuous Implementation," *Review of Economic Studies* **56**, 603-611.

Reny, P. J. (2001), "Arrow's Theorem and the Gibbard-Satterthwaite Theorem: a Unified Approach," *Economics Letters*, 99-105.

Repullo, R. (1987), "A Simple Proof of Maskin Theorem on Nash Implementation," *Social Choice and Welfare* **4**, 39-41.

Roth, A. E. (2002), "The Economist as Engineer: Game Theory, Experimentation and Computation as Tools for Design Economics," *Econometrica* **70**, 1341-1378.

Rubinstein, A. (1982), "Perfect Equilibrium in a Bargaining Model," *Econometrica* **50**, 97-109.

Rubinstein, A. and A. Wolinsky (1992), "Renegotiation-Proof Implementation and Time Preferences," *American Economic Review* **82**, 600-614.

Saijo, T. (1987), "On Constant Maskin Monotonic Social Choice Functions," *Journal of Economic Theory* **42**, 382-386.

Saijo, T. (1988), "Strategy Space Reduction in Maskin's Theorem," *Econometrica* **56**, 693-700.

Samuelson, P. (1954), "The Pure Theory of Public Expenditure," *Review of Economics and Statistics* **36**, 387-389.

Sandholm, W. (2002), "Evolutionary Implementation and Congestion Pricing," *Review of Economic Studies* **69**, 667-689.

Satterthwaite, M. A. (1975), "Strategy-Proofness and Arrow's Conditions: Existence and Correspondence Theorems for Voting Procedures and Social Welfare Functions," *Journal of Economic Theory* **10**, 187-217.

Schmeidler, D. (1980), "Walrasian Analysis via Strategic Outcome Functions," *Econometrica,* **48**, 1585-1594.

Serrano, R. (1997), "A Comment on the Nash Program and the Theory of Implementation," *Economics Letters* **55**, 203-208.

Serrano, R. and R. Vohra (2001), "Some Limitations of Virtual Bayesian Implementation," *Econometrica* **69**, 785-792.

Serrano, R. and R. Vohra (2002), "A Characterization of Virtual Bayesian Implementation," working paper 2002-11, Department of Economics, Brown University.

Sjostrom, T. (1993), "Implementation in Perfect Equilibria," *Social Choice and Welfare* **10**, 97-106.

Sjostrom, T. (1994), "Implementation in Undominated Nash Equilibria without Using Integer Games," *Games and Economic Behavior* **6**, 502-511.

Smith, A. (1776), *An Inquiry into the Nature and Causes of the Wealth of Nations*, Edinburgh.

Sprumont, Y. (1991), "A Division Problem with Single-Peaked Preferences: a Characterization of the Uniform Allocation Rules," *Econometrica* **59**, 509-519.

Tian, G. (1989), "Implementation of the Lindahl Correspondence by a Single-Valued, Feasible and continuous Mechanism," *Review of Economic Studies* **56**, 613-621.

Tian, G. (1993), "Implementing Lindahl Allocations by a Withholding Mechanism," *Journal of Mathematical Economics* **22**, 169-179.

Tian, G. (1994), "Implementation of Linear Cost Share Equilibrium Allocations," *Journal of Economic Theory* **64**, 568-584.

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

von Neumann, J. and O. Morgenstern (1944), *Theory of Games and Economic Behavior*, Princeton University Press.

Walker, M. (1981), "A Simple Incentive Compatible Mechanism for Attaining Lindhal Allocations," *Econometrica* **49**, 65-73.

Williams, S. (1986), "Realization of Nash Implementation: Two Aspects of Mechanism Design," *Econometrica* **54**, 139-151.

Yamato, T. (1993), "Double Implementation in Nash and Undominated Nash Equilibrium," *Journal of Economic Theory* **59**, 311-323.

Young, H. P. (1998), *Individual Strategy and Social Structure: an Evolutionary Theory of Institutions*, Princeton university Press.