Chapter 5

IMPLEMENTATION THEORY*

ERIC MASKIN

Institute for Advanced Study, Princeton, NJ, USA

TOMAS SJÖSTRÖM

Department of Economics, Pennsylvania State University, University Park, PA, USA

Contents

* We are grateful to Sandeep Baliga, Luis Corchón, Matt Jackson, Byungchae Rhee, Ariel Rubinstein, Ilya Segal, Hannu Vartiainen, Masahiro Watabe, and two referees, for helpful comments.

Handbook of Social Choice and Welfare, Volume 1, Edited by K.J. Arrow, A.K. Sen and K. Suzumura

Abstract

The implementation problem is the problem of designing a mechanism (game form) such that the equilibrium outcomes satisfy a criterion of social optimality embodied in a social choice rule. If a mechanism has the property that, in each possible state of the world, the set of equilibrium outcomes equals the set of optimal outcomes identified by the social choice rule, then the social choice rule is said to be implemented by this mechanism. Whether or not a social choice rule is implementable may depend on which game-theoretic solution concept is used. The most demanding requirement is that each agent should always have a dominant strategy, but mainly negative results are obtained in this case. More positive results are obtained using less demanding solution concepts such as Nash equilibrium. Any Nash-implementable social choice rule must satisfy a condition of "monotonicity". Conversely, any social choice rule which satisfies monotonicity and "no veto power" can be Nash-implemented. Even non-monotonic social choice rules can be implemented using Nash equilibrium refinements. The implementation problem can be made more challenging by imposing additional requirements on the mechanisms, such as robustness to renegotiation and collusion. If the agents are incompletely informed about the state of the world, then the concept of Nash equilibrium is replaced by Bayesian Nash equilibrium. Incentive compatibility is a necessary condition for Bayesian Nash implementation, but in other respects the results closely mimic those that obtain with complete information.

Keywords

social choice, implementation, mechanism design

JEL classification: D71

1. Introduction

The problem of social decision making when information is decentralized has occupied economists since the days of Adam Smith. An influential article by Hayek crystallized the problem. Since "the data from which the economic calculus starts are never for the whole society given to a single mind", the problem to be solved is "how to secure the best use of resources known to any of the members of society, for ends whose relative importance only these individuals know" [Hayek (1945)]. A resource allocation mechanism is thus essentially a system for communicating and processing information. A mathematical analysis of these issues became possible after the contributions of Leo Hurwicz. Hurwicz (1960, 1972) provided a formal definition of a resource allocation mechanism that is so general that almost any conceivable method for making social decisions is a possible mechanism in this framework. Hurwicz (1972) also introduced the fundamental notion of incentive compatibility.

The theory of mechanism design provides an analytical framework for the design of institutions, with emphasis on the problem of incentives[1]. A mechanism, or game form, is thought of as specifying the rules of a game. The players are the members of the society (the agents). The question is whether the equilibrium outcomes will be, in some sense, socially optimal. Formally, the problem is formulated in terms of the implementation of social choice rules. A social choice rule specifies, for each possible state of the world, which outcomes would be socially optimal in that state. It can be thought of as embodying the welfare judgements of a social planner. Since the planner does not know the true state of the world, she must rely on the agents' equilibrium actions to indirectly cause the socially optimal outcome to come about. If a mechanism has the property that, in each possible state of the world, the set of equilibrium outcomes equals the set of socially optimal outcomes identified by the social choice rule, then the social choice rule is said to be implemented by this mechanism. By definition, implementation is easier to accomplish the smaller is the set of possible states of the world. For example, if the social planner knows that each agent's true utility function belongs to the class of quasi-linear utility functions, then her task is likely to be simpler than if she had no such prior information.

To be specific, consider two kinds of decision problems a society may face. The first is the economic problem of producing and allocating private and/or public goods. Here, a state of the world specifies the preferences, endowments, and productive technology of each economic agent (normally, certain a priori restrictions are imposed on the preferences, e.g., non-satiation). For economies with only private goods, traditional economic theory has illuminated the properties of the competitive price system. In our terminology, the Walrasian rule is the social choice rule that assigns to each state of the world the corresponding set of competitive (Walrasian) allocations. A mechanism

[1] Other surveys that cover much of the material we discuss here include Maskin (1985), Groves and Ledyard (1987), Moore (1992), Palfrey (1992, 2001), Corchón (1996) and Jackson (2001).

might involve agents announcing prices and quantities, or perhaps only quantities (the appropriate prices could be calculated by a computer). To solve the implementation problem we need to verify that the set of equilibrium outcomes of the mechanism coincides with the set of Walrasian allocations in each possible state of the world. In public goods economies, we may instead be interested in implementing the Lindahl rule, i.e., the social choice rule that assigns to each state of the world its corresponding set of Lindahl allocations (these are the competitive equilibrium allocations in the fictitious price system where each consumer has a personalized price for each public good). Of course, the Walrasian and Lindahl rules are only two examples of social choice rules in economic environments. More generally, implementation theory characterizes the full class of implementable social choice rules.

A second example of a social decision problem is the problem of choosing one alternative from a finite set (e.g., selecting a president from a set of candidates). In this environment, a social choice rule is often called a voting rule. No restrictions are necessarily imposed on how the voters may rank the alternatives. When the feasible set consists of only two alternatives, then a natural voting rule is the ordinary method of majority rule. But with three or more alternatives, there are many plausible voting rules, such as Borda's rule[2] and other rank-order voting schemes. Again, implementation theory characterizes the set of implementable voting rules.

Whether or not a social choice rule is implementable may depend on which game theoretic solution concept is invoked. The most demanding requirement is that each agent should have a dominant strategy. A mechanism with this property is called a dominant strategy mechanism. By definition, a dominant strategy is optimal for the agent regardless of the actions of others. Thus, in a dominant strategy mechanism agents need not form any conjecture about the behavior of others in order to know what to do. The revelation principle, first stated by Gibbard (1973), implies that there is a sense in which the search for dominant strategy mechanisms may be restricted to "revelation mechanisms" in which each agent simply reports his own personal characteristics (preferences, endowments, productive capacity ...) to the social planner. The planner uses this information to compute the state of the world and then chooses the outcome that the social choice rule prescribes in this state. (To avoid the difficulties caused by tie-breaking, assume the social choice rule is single-valued.) Of course, the chosen outcome is unlikely to be socially optimal if agents misrepresent their characteristics. A social choice rule is dominant strategy incentive compatible, or strategy-proof, if the associated revelation mechanism has the property that honestly reporting the truth is always a dominant strategy for each agent.

Unfortunately, in many environments no satisfactory strategy-proof social choice rules exist. For the classical private goods economy, Hurwicz (1972) proved that no

[2] If there are m alternatives, then Borda's rule assigns each alternative m points for every agent who ranks it first, $m-1$ points for every agent who ranks it second, etc.; the winner is the alternative with the biggest point total.

Pareto optimal and individually rational social choice rule can be strategy-proof if the space of admissible preferences is large enough[3]. An analogous result was obtained for the classical public goods economy by Ledyard and Roberts (1974). It follows from these results that neither the Walrasian rule nor the Lindahl rule is strategy-proof. These results confirmed the suspicions of many economists. In particular, Vickrey (1961) conjectured that if an agent was not negligibly small compared to the whole economy, then any attempt to allocate divisible private goods in a Pareto optimal way would imply "a direct incentive for misrepresentation of the marginal-cost or marginal-value curves". Samuelson (1954) argued that no resource allocation mechanism could generate a Pareto optimal level of public goods because "it is in the selfish interest of each person to give *false* signals, to pretend to have less interest in a given collective activity than he really has, etc"[4].

If only quasi-linear utility functions are admissible (utility functions are additively separable between the public decision and money and linear in money), then there does exist an attractive class of mechanisms, the Vickrey–Groves–Clarke mechanisms, with the property that truth-telling is a dominant strategy [Vickrey (1961), Groves (1970), Clarke (1971)]. But a Vickrey–Groves–Clarke mechanism will in general fail to balance the budget (the monetary transfers employed to induce truthful revelation do not sum to zero), and so Vickrey's and Samuelson's pessimistic conjectures were formally correct even in the quasi-linear case [Green and Laffont (1979), Walker (1980), Hurwicz and Walker (1990)][5].

The search for dominant strategy mechanisms in the case of voting over a finite set of alternatives turned up even more negative results. Gibbard (1973) and Satterthwaite (1975) showed that if the range of a strategy-proof voting rule contains at least three alternatives then it must be dictatorial, assuming the set of admissible preferences contains all strict orderings. Again, this impossibility result confirmed the suspicions of many economists, notably Arrow (1963), Vickrey (1960) and Dummett and Farquharson (1961). It follows that the Borda rule, for example, is not strategy-proof. In fact, Borda himself knew that his scheme was vulnerable to insincere voting and had intended it to be used only by "honest men" [Black (1958)].

If we drop the requirement that each agent should have a dominant strategy then the situation is much less bleak. The idea of Nash equilibrium is fundamental to much of economic theory. In a Nash equilibrium, each agent's action is a best response to the actions that he predicts other agents will take, and in addition these predictions are correct. Formal justifications of this concept usually rely on each agent having complete information about the state of the world. If agents have complete information

[3] Hurwicz's (1972) definition of incentive compatibility was essentially a requirement that truthful reports should be a Nash equilibrium in a game where each agent reports his own personal characteristics (at a minimum, an agent's "personal characteristics" determine his preferences). This implies that truth-telling is a dominant strategy.

[4] An early discussion of the incentives to manipulate the Lindahl rule can be found in Bowen (1943).

[5] But see Groves and Loeb (1975) for a special quadratic case where budget balance is possible.

in this sense, then the planner can ask each agent to report the complete state of the world, not just his own characteristics[6]. With at least three agents, and with the planner disregarding a single dissenting opinion against a consensus, it is a Nash equilibrium for all agents to announce the state truthfully (each agent is using a best response because he cannot change the outcome by deviating unilaterally). However, this kind of revelation mechanism would also have many non-truthful Nash equilibria. This highlights a general difficulty with the revelation principle: although incentive compatibility guarantees that truth-telling is an equilibrium, it does not guarantee that it is the *only* equilibrium. The implementation literature normally requires that *all* equilibrium outcomes should be socially optimal (an exception is the dominant-strategy literature, where the possibility of multiple equilibria, i.e., multiple dominant strategies, is typically much less worrisome).

Nash implementation using mechanisms with general message spaces was first studied by Groves and Ledyard (1977), Hurwicz and Schmeidler (1978) and Maskin (1999)[7]. For a class of economic environments, Groves and Ledyard (1977) discovered that non-dictatorial mechanisms exist such that all Nash equilibrium outcomes are Pareto optimal. Hurwicz and Schmeidler (1978) found a similar result for the case of social choice from a finite set of alternatives. General results applicable to both kinds of environments were obtained by Maskin (1999). He found that a "monotonicity" condition is necessary for a social choice rule to be Nash-implementable. With at least three agents, monotonicity plus a condition of "no veto power" is sufficient. The monotonicity condition says that if a socially optimal alternative does not fall in any agent's preference ordering relative to any other alternative, then it remains socially optimal. In economic environments, the Walrasian and Lindahl rules satisfy monotonicity (strictly speaking, the Walrasian and Lindahl rules have to be modified slightly to render them monotonic). Since no veto power is always satisfied in economic environments with three or more non-satiated agents, these social choice rules can be Nash-implemented. In the case of voting with a finite set of alternatives, a monotonic *single-valued* social choice rule must be dictatorial if the preference domain consists of all strict orderings, and there are (at least) three different alternatives such that for each of them there is a state where that alternative is socially optimal. However, the (weak) Pareto correspondence is a monotonic social choice *correspondence* that satisfies no veto power in any environment, and hence it can be Nash-implemented.

[6] Such a mechanism requires transmission of an enormous amount of information to the social planner. In practice, this may be costly and time-consuming. However, in this survey we do not focus on the issue of informational efficiency, but rather on characterization of the set of implementable social choice rules. The mechanisms are not intended to be "realistic", and in applications one would look for much simpler mechanisms. It is worth noticing that in Hurwicz's (1960) original "decentralized mechanism", messages were simply sets of net trade vectors. Important theorems concerning the informational efficiency of price mechanisms were established by Mount and Reiter (1974) and Hurwicz (1977).

[7] Maskin's article was circulated as a working paper in 1977.

If agent i's strategy s_i is a best response against the strategies of others, and the resulting outcome is a, then s_i remains a best response if outcome a moves up in agent i's preference ordering. Thus, such a change in agent i's preferences cannot destroy a Nash equilibrium (which is why monotonicity is a necessary condition for Nash implementation). However, it can make s_i a weakly dominated strategy for agent i, and so can destroy an undominated Nash equilibrium (i.e., a Nash equilibrium where each agent is using a weakly undominated strategy). Hence monotonicity is not a necessary condition for implementation in undominated Nash equilibria.

This insight was exploited by Palfrey and Srivastava (1991), who found that many more social choice rules can be implemented in undominated Nash equilibria than in Nash equilibria. A similar result was found by Sjöström (1993) for implementation in trembling-hand perfect Nash equilibria[8]. Moreover, rather different paths can lead to the implementation of non-monotonic social choice rules. Moore and Repullo (1988) showed that the set of implementable social choice rules can be dramatically expanded by the use of extensive game forms. This development was preceded by the work by Farquharson (1969) and Moulin (1979) on sequential voting mechanisms. Abreu and Sen (1991) and Matsushima (1988) considered "virtual" implementation, where the socially optimal outcome is required to occur only with probability close to one, and found that the set of virtually implementable social choice rules is also very large.

Despite this plethora of positive results, it would not be correct to say that *any* social choice rule can be implemented by a sufficiently clever mechanism together with a suitable refinement of Nash equilibrium. Specifically, only *ordinal* social choice rules can be implemented[9]. This is a significant restriction since many well-known social welfare criteria depend on cardinal information about preferences (for example, utilitarianism and various forms of egalitarianism). On the other hand, if there are at least three agents, then, with suitable equilibrium refinement, not much more than ordinality is required for implementation[10]. The mechanisms that are used to establish these most general "possibility theorems" sometimes have a questionable feature, viz., out-of-equilibrium behavior may lead to highly undesirable outcomes (for example, worthwhile goods may be destroyed). If the agents can renegotiate such bad outcomes then such mechanisms no longer work [Maskin and Moore (1999)]. In fact, the

[8] Nash equilibrium refinements help implementation by destroying undesirable equilibria, but they also make it harder to support a socially optimal outcome as an equilibrium outcome. In practice, refinements seem to help more often than they hurt, but it is not difficult to come up with counter-examples. Sjöström (1993) gives an example of a social choice rule that is implementable in Nash equilibria but not in trembling-hand perfect Nash equilibria.

[9] An ordinal social choice rule does not rely on cardinal information about the "intensity" of preference. Thus, if the social choice rule prescribes different outcomes in two different states, then there must exist some agent i and some outcomes a and b such that agent i's ranking of a versus b is not the same in the two states (i.e., there is preference reversal).

[10] Sometimes the no veto power condition is part of the sufficient condition. Although no veto power is normally trivially satisfied in economic environments with at least three agents, it is not always an innocuous condition in other environments.

possibility of renegotiation can make the implementation problem significantly more difficult when there are only two agents. However, the general "possibility theorems" seem to survive renegotiation in economic environments with three or more agents [Sjöström (1999)].

Obviously, the social planner cannot freely "choose" a solution concept (such as undominated Nash equilibrium) to suit his purposes. In some sense, the solution concept should be appropriate for the mechanism and environment at hand, but it is hard to make this requirement mathematically precise [for an insightful discussion, see Jackson (1992)]. Harsanyi and Selten (1988) argue that game theoretic analysis should lead to an ideal solution concept that applies universally to all possible games, but experiments show that behavior in practice depends on the nature of the game (even on "irrelevant" aspects such as the labelling of strategies). How the mechanism is explained to the agents may be an important part of the design process (e.g., "please notice that strategy s_i is dominated"). Hurwicz (1972) argued in terms of a dynamic adjustment toward Nash equilibrium: each agent would keep modifying his strategy according to a fixed "response function" until a Nash equilibrium was reached. However, Jordan (1986) showed that equilibria of game forms that Nash-implement the Walrasian rule will in general not be stable under continuous-time strategy-adjustment processes. Muench and Walker (1984), de Trenqualye (1988) and Cabrales (1999) also discuss the problem of how agents may come to coordinate on a particular equilibrium. Cabrales and Ponti (2000) show how evolutionary dynamics may lead to the "wrong" Nash equilibrium in mechanisms which rely on the elimination of weakly dominated strategies. Best-response dynamics do converge to the "right" equilibrium in the particular mechanism they analyze. But these kinds of naive adjustment processes are difficult to interpret, because behavior is not fully rational along the path: a fully rational agent would try to exploit the naivete of other agents, especially if he knew (or could infer something about) their payoff functions. In experiments where a game is played repeatedly, treatments in which players are uninformed about the payoff functions of other players appear more likely to end up at a Nash equilibrium (of the one-shot game) than treatments where players do have this information [Smith (1979)]. Perhaps it is too difficult to even attempt to manipulate the behavior of an opponent with an unknown payoff function. It was precisely because he did not want to assume that agents have complete information that Hurwicz (1972) introduced the dynamic adjustment processes. But the problem of how agents can learn to play a Nash equilibrium is difficult [for a good introduction, see Fudenberg and Levine (1998)].

If we discount the possibility that incompletely informed agents will end up at a Nash equilibrium, then the results of Maskin (1999) and the literature that followed him can be interpreted as drawing out the logical implications of the assumption that agents have complete information about the state of the world. In some cases this assumption may be reasonable, and many economic models explicitly or implicitly rely on it. But in other cases it makes more sense to assume that agents assign positive

probability to many different states of the world, and behave as Bayesian expected utility maximizers.

Bayesian mechanism design was pioneered by D'Aspremont and Gérard-Varet (1979), Dasgupta, Hammond and Maskin (1979), Myerson (1979) and Harris and Townsend (1981). If an agent has private information not shared by other agents, then a Bayesian incentive compatibility condition is necessary for him to be willing to reveal it. But not every Bayesian incentive compatible social choice rule is Bayesian Nash-implementable, because a revelation mechanism may have undesirable equilibria in addition to the truthful one. Postlewaite and Schmeidler (1986), Palfrey and Srivastava (1989a) and Jackson (1991) have shown that the results of Maskin (1999) can be generalized to the Bayesian environment. A Bayesian monotonicity condition is necessary for Bayesian Nash implementation. With at least three agents, a condition that combines Bayesian monotonicity with no veto power is sufficient for implementation, as long as Bayesian incentive compatibility and a necessary condition called closure are satisfied [Jackson (1991)].

Mechanisms can also be used to represent *rights* [Gärdenfors (1981), Gaertner, Pattanaik and Suzumura (1992), Deb (1994), Hammond (1997)]. Deb, Pattanaik and Razzolini (1997) introduced several properties of mechanisms that correspond to "acceptable" rights structures. For example, an individual has a *say* if there exists at least some circumstance where his actions can influence the outcome[11]. The notion of rights is important but will not be discussed in this survey. Our notion of implementation is consequentialist: the precise structure of a mechanism does not matter as long as its equilibrium outcomes are socially optimal.

2. Definitions

The *environment* is $\langle A, N, \Theta \rangle$, where A is the set of feasible *alternatives* or *outcomes*, $N = \{1, 2, \ldots, n\}$ is the finite set of *agents*, and Θ is the set of possible *states of the world*. For simplicity, we suppose that the set of feasible alternatives is the same in all states [see Hurwicz, Maskin and Postlewaite (1995) for implementation with a state-dependent feasible set]. The agents' preferences do depend on the state of the world. Each agent $i \in N$ has a payoff function u_i: $A \times \Theta \to \boldsymbol{R}$. Thus, if the outcome is $a \in A$ in state of the world $\theta \in \Theta$, then agent i's payoff is $u_i(a, \theta)$. His weak preference relation in state θ is denoted $R_i = R_i(\theta)$, the strict part of his preference is denoted $P_i = P_i(\theta)$, and indifference is denoted $I_i = I_i(\theta)$. That is, $xR_i y$ if and only if $u_i(x, \theta) \geqslant u_i(y, \theta)$, $xP_i y$ if and only if $u_i(x, \theta) > u_i(y, \theta)$, and $xI_i y$ if and only if $u_i(x, \theta) = u_i(y, \theta)$. The *preference profile* in state $\theta \in \Theta$ is denoted

[11] Gaspart (1996, 1997) proposed a stronger notion of equality (or symmetry) of attainable sets: all agents, by unilaterally varying their actions, should be able to attain identical (or symmetric) sets of outcomes, at least at equilibrium.

$R = R(\theta) = (R_1(\theta), \ldots, R_n(\theta))$. The *preference domain* is the set of preference profiles that are consistent with some state of the world, i.e., the set

$$\mathcal{R}(\Theta) \equiv \{R\text{: there is } \theta \in \Theta \text{ such that } R = R(\theta)\}.$$

The *preference domain for agent i* is the set

$$\mathcal{R}_i(\Theta) \equiv \{R_i\text{: there is } R_{-i} \text{ such that } (R_i, R_{-i}) \in \mathcal{R}(\Theta)\}.$$

When Θ is fixed, we can write $\mathcal{R}$ and $\mathcal{R}_i$ instead of $\mathcal{R}(\Theta)$ and $\mathcal{R}_i(\Theta)$.

Let $\mathcal{R}_A$ be the set of *all* profiles of complete and transitive preference relations on A, the *unrestricted domain*. It will always be true that $\mathcal{R}(\Theta) \subseteq \mathcal{R}_A$. Let $\mathcal{P}_A$ be the set of all profiles of linear orderings of A, the *unrestricted domain of strict preferences*[12].

For any sets X and Y, let $X - Y \equiv \{x \in X\text{: } x \notin Y\}$, let Y^X denote the set of all functions from X to Y, and let 2^X denote the set of all subsets of X. If X is finite, then $|X|$ denotes the number of elements in X.

A *social choice rule* (SCR) is a function $F\text{: } \Theta \to 2^A - \{\emptyset\}$ (i.e., a non-empty valued correspondence). The set $F(\theta) \subseteq A$ is the set of socially optimal (or *F-optimal*) alternatives in state $\theta \in \Theta$. The *image* or *range* of the SCR F is the set

$$F(\Theta) \equiv \{a \in A\text{: } a \in F(\theta) \text{ for some } \theta \in \Theta\}.$$

A *social choice function* (SCF) is a *single-valued* SCR, i.e., a function $f\text{: } \Theta \to A$.

Some important properties of SCRs are as follows.

- *Ordinality*: for all $(\theta, \theta') \in \Theta \times \Theta$, if $R(\theta) = R(\theta')$ then $F(\theta) = F(\theta')$.
- *Weak Pareto optimality*: for all $\theta \in \Theta$ and all $a \in F(\theta)$, there is no $b \in A$ such that $u_i(b, \theta) > u_i(a, \theta)$ for all $i \in N$.
- *Pareto optimality*: for all $\theta \in \Theta$ and all $a \in F(\theta)$, there is no $b \in A$ such that $u_i(b, \theta) \geqslant u_i(a, \theta)$ for all $i \in N$ with strict inequality for some i.
- *Pareto indifference*: for all $(a, \theta) \in A \times \Theta$ and all $b \in F(\theta)$, if $u_i(a, \theta) = u_i(b, \theta)$ for all $i \in N$ then $a \in F(\theta)$.
- *Dictatorship*: there exists $i \in N$ such that for all $\theta \in \Theta$ and all $a \in F(\theta)$, $u_i(a, \theta) \geqslant u_i(b, \theta)$ for all $b \in A$.
- *Unanimity*: for all $(a, \theta) \in A \times \Theta$, if $u_i(a, \theta) \geqslant u_i(b, \theta)$ for all $i \in N$ and all $b \in A$ then $a \in F(\theta)$.
- *Strong unanimity*: for all $(a, \theta) \in A \times \Theta$, if $u_i(a, \theta) > u_i(b, \theta)$ for all $i \in N$ and all $b \neq a$ then $F(\theta) = \{a\}$.

[12] A preference relation R_i is a linear ordering if and only if it is complete, transitive and antisymmetric (for all $(a, b) \in A \times A$, if aR_ib and bR_ia then $a = b$).

– *No veto power*: for all $(a,j,\theta) \in A \times N \times \Theta$, if $u_i(a,\theta) \geqslant u_i(b,\theta)$ for all $b \in A$ and all $i \neq j$ then $a \in F(\theta)$.

A *mechanism* (or *game form*) is denoted $\Gamma = \langle \times_{i=1}^n M_i, h \rangle$ and consists of a *message space* M_i for each agent $i \in N$ and an *outcome function* h: $\times_{i=1}^n M_i \to A$. Let $m_i \in M_i$ denote agent i's message. A *message profile* is denoted $m = (m_1, \ldots, m_n) \in M \equiv \times_{i=1}^n M_i$. All messages are sent simultaneously, and the final outcome is $h(m) \in A$. This kind of mechanism is sometimes called a *normal form mechanism* (or *normal game form*) to distinguish it from *extensive form mechanisms* in which agents make choices sequentially [Moore and Repullo (1988)]. With the exception of Section 4.4, nearly all our results relate to normal form mechanisms, so merely calling them "mechanisms" should not cause confusion.

The most common interpretation of the implementation problem is that a *social planner* or *mechanism designer* (who cannot observe the true state of the world) wants to design a mechanism in such a way that in each state of the world the set of equilibrium outcomes coincides with the set of F-optimal outcomes. Let $\mathcal{S}$ equilibrium be a game theoretic solution concept and let F be an SCR. For each mechanism Γ and each state $\theta \in \Theta$, the solution concept specifies a set of $\mathcal{S}$ *equilibrium outcomes* denoted $\mathcal{S}(\Gamma,\theta) \subseteq A$. A mechanism Γ *implements* F *in* $\mathcal{S}$ *equilibria*, or simply $\mathcal{S}$*-implements* F, if and only if $\mathcal{S}(\Gamma,\theta) = F(\theta)$ for all $\theta \in \Theta$. Thus, the set of $\mathcal{S}$ equilibrium outcomes should coincide with the set of F-optimal outcomes in each state. If such a mechanism exists then F is *implementable in* $\mathcal{S}$ *equilibria* or simply $\mathcal{S}$*-implementable*. This notion is sometimes referred to as *full implementation*. Clearly, whether or not an SCR F is $\mathcal{S}$-implementable may depend on the solution concept $\mathcal{S}$. If solution concept $\mathcal{S}_2$ is a refinement of $\mathcal{S}_1$, in the sense that for any Γ we have $\mathcal{S}_2(\Gamma,\theta) \subseteq \mathcal{S}_1(\Gamma,\theta)$ for all $\theta \in \Theta$, then it is not a priori clear whether it will be easier to satisfy $\mathcal{S}_1(\Gamma,\theta) = F(\theta)$ or $\mathcal{S}_2(\Gamma,\theta) = F(\theta)$ for all $\theta \in \Theta$. However, as discussed in the Introduction, the literature shows that refinements "usually" make things easier.

Most of this survey deals with full implementation in the above sense, but we will briefly deal with the notions of *weak* and *double* implementation. A mechanism Γ *weakly* $\mathcal{S}$*-implements* F if and only if $\emptyset \neq \mathcal{S}(\Gamma,\theta) \subseteq F(\theta)$ for all $\theta \in \Theta$. That is, every $\mathcal{S}$ equilibrium outcome must be F-optimal, but every F-optimal outcome need not be an equilibrium outcome. Weak implementation is actually subsumed by the theory of full implementation, since weak implementation of F is equivalent to full implementation of a subcorrespondence of F [Thomson (1996)]. If $\mathcal{S}_1$ and $\mathcal{S}_2$ are two solution concepts, then Γ *doubly* $\mathcal{S}_1$*- and* $\mathcal{S}_2$*-implements* F if and only if $\mathcal{S}_1(\Gamma,\theta) = \mathcal{S}_2(\Gamma,\theta) = F(\theta)$ for all $\theta \in \Theta$.

3. Nash implementation

We start by assuming that the true state of the world is common knowledge among the agents. This is the case of *complete information*. We will consider mechanisms in normal form. (Extensive form mechanisms are discussed in Section 4.4.)

3.1. Definitions

Given a mechanism $\Gamma = \langle M, h \rangle$ for any $m \in M$ and $i \in N$, let $m_{-i} = \{m_j\}_{j \neq i} \in M_{-i} \equiv \times_{j \neq i} M_j$ denote the messages sent by agents other than i. For message profile $m = (m_{-i}, m_i) \in M$, the set

$$h(m_{-i}, M_i) \equiv \{a \in A\colon a = h(m_{-i}, m_i') \text{ for some } m_i' \in M_i\}$$

is agent i's *attainable set at* m. Agent i's *lower contour set* at $(a, \theta) \in A \times \Theta$ is $L_i(a, \theta) \equiv \{b \in A\colon u_i(a, \theta) \geqslant u_i(b, \theta)\}$. A message profile $m \in M$ is a (pure strategy) *Nash equilibrium* at state $\theta \in \Theta$ if and only if $h(m_{-i}, M_i) \subseteq L_i(h(m), \theta)$ for all $i \in N$. (For now we neglect mixed strategies: they are discussed in Section 4.3.) The set of Nash equilibria at state θ is denoted $N^\Gamma(\theta) \subseteq M$, and the set of *Nash equilibrium outcomes* at state θ is denoted $h(N^\Gamma(\theta)) = \{a \in A\colon a = h(m) \text{ for some } m \in N^\Gamma(\theta)\}$. The mechanism Γ *Nash-implements* F if and only if $h(N^\Gamma(\theta)) = F(\theta)$ for all $\theta \in \Theta$.

3.2. Monotonicity and no veto power

If $L_i(a, \theta) \subseteq L_i(a, \theta')$ then we say that $R_i(\theta')$ is a *monotonic transformation* of $R_i(\theta)$ *at alternative* a. The SCR F is *monotonic* if and only if for all $(a, \theta, \theta') \in A \times \Theta \times \Theta$ the following is true: if $a \in F(\theta)$ and $L_i(a, \theta) \subseteq L_i(a, \theta')$ for all $i \in N$, then $a \in F(\theta')$. Thus, monotonicity requires that if a is optimal in state θ, and when the state changes from θ to θ' outcome a does not fall in any agent's preference ordering relative to any other alternative, then a remains optimal in state θ'. Clearly, if F is monotonic then it must be ordinal. But many ordinal social choice rules are not monotonic[13]. Whether a particular SCR is monotonic may depend on the preference domain $\mathcal{R}(\Theta)$. For example, in an exchange economy, the Walrasian correspondence is not monotonic in general, but it is monotonic on a domain of preferences such that all Walrasian equilibria occur in the interior of the feasible set [Hurwicz, Maskin and Postlewaite (1995)]. There is no monotonic and Pareto optimal SCR on the unrestricted domain $\mathcal{R}_A$ [Hurwicz and Schmeidler (1978)][14]. However, the weak Pareto correspondence[15] is monotonic on any domain. A monotonic SCF on $\mathcal{R}_A$ must be a constant function[16], but there are important examples of monotonic non-constant SCFs on restricted domains.

[13] If F is not monotonic then an interesting problem is to find the *minimal monotonic extension*, i.e., the smallest monotonic supercorrespondence of F [Sen (1995), Thomson (1999)].

[14] Let $\theta \in \Theta$ be a state where the agents do not unanimously agree on a top-ranked alternative, and let $a \in F(\theta)$. There must exist $j \in N$ and $b \in A$ such that $bP_j(\theta)a$. Let state θ' be such that preferences over alternatives in $A - \{b\}$ are as in state θ, but each agent $i \neq j$ has now become indifferent between a and b. Agent j still strictly prefers b to a in state θ' so b Pareto dominates a. But $L_i(a, \theta) \subseteq L_i(a, \theta')$ for all i so $a \in F(\theta')$ if F is monotonic, a contradiction of Pareto optimality.

[15] The weak Pareto correspondence selects all weakly Pareto optimal outcomes: for all $\theta \in \Theta$, $F(\theta) = \{a \in A\colon$ there is no $b \in A$ such that $u_i(b, \theta) > u_i(a, \theta)$ for all $i \in N\}$.

[16] That is, $f(\Theta) = \{a\}$ for some $a \in A$. For if $f(\theta) = a \neq a' = f(\theta')$ then monotonicity implies $\{a, a'\} \subseteq f(\theta'')$ if a and a' are both top-ranked by all agents in state θ'', but this contradicts the fact that f is single-valued. See Saijo (1987).

Maskin (1999) proved that for any mechanism Γ, the Nash equilibrium outcome correspondence $h \circ N^{\Gamma}: \Theta \to A$ is monotonic.

Theorem 1: [Maskin (1999)]. *If the SCR F is Nash-implementable, then F is monotonic.*

Proof: Suppose $\Gamma = \langle M, h \rangle$ Nash-implements F. Then if $a \in F(\theta)$ there is $m \in N^{\Gamma}(\theta)$ such that $a = h(m)$. Suppose $L_i(a, \theta) \subseteq L_i(a, \theta')$ for all $i \in N$. Then, for all $i \in N$,

$$h(m_{-i}, M_i) \subseteq L_i(a, \theta) \subseteq L_i(a, \theta').$$

Therefore, $m \in N^{\Gamma}(\theta')$, and so $a \in h(N^{\Gamma}(\theta')) = F(\theta')$. □

Theorem 1 has a partial converse. It was originally stated by Maskin in 1977, but without a complete proof [see Maskin (1999)]. Rigorous proofs were given by Williams (1986), Repullo (1987) and Saijo (1988). Recall that F satisfies no veto power if an alternative is F-optimal whenever it is top-ranked by at least $n - 1$ agents. In economic environments, no veto power is usually vacuously satisfied (because two different agents will never share the same top-ranked alternative). However, in other environments no veto power may not be a trivial condition. If, for example, A is a finite set, $\mathcal{R}(\Theta) = \mathcal{P}_A$ and the number of alternatives is strictly greater than the number of agents, then even the Borda rule does not satisfy no veto power[17]. If $\mathcal{R}(\Theta) = \mathcal{R}_A$ then no Pareto optimal SCR can satisfy no veto power[18]. Still, the weak Pareto correspondence satisfies no veto power on any domain.

Theorem 2: [Maskin (1999)]. *Suppose $n \geqslant 3$. If the SCR F satisfies monotonicity and no veto power, then F is Nash-implementable.*

Proof: The proof is constructive. Let each agent $i \in N$ announce an outcome, a state of the world, and an integer between 1 and n. Thus, $M_i = A \times \Theta \times \{1, 2, \ldots, n\}$ and a typical message for agent i is denoted $m_i = (a^i, \theta^i, z^i) \in M_i$. Let the outcome function be as follows.

Rule 1: If $(a^i, \theta^i) = (a, \theta)$ for all $i \in N$ and $a \in F(\theta)$, then $h(m) = a$.

Rule 2: Suppose there exists $j \in N$ such that $(a^i, \theta^i) = (a, \theta)$ for all $i \neq j$ but $(a^j, \theta^j) \neq (a, \theta)$. Then $h(m) = a^j$ if $a^j \in L_j(a, \theta)$ and $h(m) = a$ otherwise.

Rule 3: In all other cases, let $h(m) = a^j$ for $j \in N$ such that $j = (\sum_{i \in N} z^i) \pmod n$ [19].

We need to show that, for any $\theta^* \in \Theta$, $h(N^{\Gamma}(\theta^*)) = F(\theta^*)$.

Step 1: $h(N^{\Gamma}(\theta^*)) \subseteq F(\theta^*)$. Suppose $m \in N^{\Gamma}(\theta^*)$. If either rule 2 or rule 3 applies to m, then there is $j \in N$ such that any agent $k \neq j$ can get his top-ranked alternative,

[17] Suppose agent 1 ranks a first and b last. All other agents rank b first and a second. If $|A| > n$ then b gets a lower Borda score than a and hence is not selected.

[18] If $u_1(b, \theta) > u_1(a, \theta)$, and $u_i(b, \theta) = u_i(a, \theta) \geqslant u_i(x, \theta)$ for all $i \neq 1$ and all $x \in A - \{a, b\}$, then no veto power implies $a \in F(\theta)$ even though b Pareto dominates a.

[19] $\alpha = \beta \pmod n$ denotes that integers α and β are congruent modulo n.

via rule 3, by announcing an integer z^k such that $k = (\sum z^i) \pmod n$. Therefore, we must have $u_k(h(m), \theta^*) \geqslant u_k(x, \theta^*)$ for all $k \neq j$ and all $x \in A$, and hence $h(m) \in F(\theta^*)$ by no veto power. If instead rule 1 applies, then $(a^i, \theta^i) = (a, \theta)$ for all $i \in N$, and $a \in F(\theta)$. The attainable set for each agent j is $L_j(a, \theta)$ by rule 2. Since $m \in N^\Gamma(\theta^*)$, we have $L_j(a, \theta) \subseteq L_j(a, \theta^*)$. By monotonicity, $a \in F(\theta^*)$. Thus, $h(N^\Gamma(\theta^*)) \subseteq F(\theta^*)$.

Step 2: $F(\theta^*) \subseteq h(N^\Gamma(\theta^*))$. Suppose $a \in F(\theta^*)$. If $m_i = (a, \theta^*, 1)$ for all $i \in N$, then $h(m) = a$. By rule 2, $h(m_{-j}, M_j) = L_j(a, \theta^*)$ for all $j \in N$, so $m \in N^\Gamma(\theta^*)$. Thus, $F(\theta^*) \subseteq h(N^\Gamma(\theta^*))$. □

The mechanism in the proof of Theorem 2 is the *canonical mechanism for Nash implementation*. Rule 3 is referred to as a "modulo game".

The canonical mechanism can be simplified in several ways even in this abstract framework. Since any Nash-implementable F is ordinal, it clearly suffices to let the agents announce a preference profile $R \in \mathcal{R}(\Theta)$ rather than a state of the world $\theta \in \Theta$. In fact, it suffices if each agent $i \in N$ announces a preference ordering for himself and one for his "neighbor" agent $i + 1$, where agents 1 and n are considered neighbors [Saijo (1988)]. Lower contour sets could be announced instead of preference orderings [McKelvey (1989)]. Much less information is needed when F is the Walrasian rule [Chakravorty (1991)].

More generally, given any message process that "computes" (or "realizes") an SCR, Williams (1986) considered the problem of embedding the message process into a mechanism which Nash-implements the SCR. If the original message process encodes information in an efficient way, then the same will be true for Williams' mechanism for Nash implementation.

3.3. *Necessary and sufficient conditions*

The no veto power condition is not necessary for Nash implementation with $n \geqslant 3$. On the other hand, monotonicity on its own is not sufficient [see Maskin (1985, 1999) for a counterexample]. The necessary and sufficient condition was given by Moore and Repullo (1990). It can be explained by considering how the canonical mechanism of Section 3.2 must be modified when no veto power is violated.

Suppose we want to Nash-implement a monotonic SCR F using some mechanism $\Gamma = \langle M, h \rangle$. Let $a \in F(\theta)$. There must exist a Nash equilibrium $m^* \in N^\Gamma(\theta)$ such that $h(m^*) = a$. Agent j's attainable set must satisfy $h(m^*_{-j}, M_j) \subseteq L_j(a, \theta)$. Alternative $c \in L_j(a, \theta)$ is an *awkward outcome for agent j in $L_j(a, \theta)$* if and only if there is $\theta' \in \Theta$ such that: (i) $L_j(a, \theta) \subseteq L_j(c, \theta')$; (ii) for each $i \neq j$, $L_i(c, \theta') = A$; (iii) $c \notin F(\theta')$. Notice that there are no awkward outcomes if F satisfies no veto power, since in that case (ii) and (iii) cannot both hold. But suppose no veto power is violated and (i), (ii) and (iii) all hold for θ' so c is awkward in $L_j(a, \theta)$. If $c \in h(m^*_{-j}, M_j)$ then there is $m_j \in M_j$ such that $h(m^*_{-j}, m_j) = c$. Then $(m^*_{-j}, m_j) \in N^\Gamma(\theta')$ since (i) implies c is the best outcome for agent j in his attainable set $h(m^*_{-j}, M_j)$ in state θ', and (ii) implies c is the best outcome in all of A for all other agents. By (iii), $c \notin F(\theta')$, so $h(N^\Gamma(\theta')) \neq F(\theta')$,

contradicting the definition of implementation. Thus, the awkward outcome c cannot be in agent j's attainable set. We must have $h(m^*_{-j}, M_j) \subseteq C_j(a, \theta)$, where $C_j(a, \theta)$ denotes the set of outcomes in $L_j(a, \theta)$ that are *not* awkward for agent j in $L_j(a, \theta)$. That is, $C_j(a, \theta) \equiv \{c \in L_j(a, \theta)$: for all $\theta' \in \Theta$, if $L_j(a, \theta) \subseteq L_j(c, \theta')$ and for each $i \neq j$, $L_i(c, \theta') = A$, then $c \in F(\theta')\}$. But if $h(m^*_{-i}, M_i) \subseteq C_i(a, \theta)$ for all $i \in N$, then for any $\theta' \in \Theta$ such that $C_i(a, \theta) \subseteq L_i(a, \theta')$ for all $i \in N$ we will have $m^* \in N^\Gamma(\theta')$, so Nash implementation requires $a = h(m^*) \in F(\theta')$. The SCR F is *strongly monotonic* if and only if for all $(a, \theta, \theta') \in A \times \Theta \times \Theta$ the following is true: if $a \in F(\theta)$ and $C_i(a, \theta) \subseteq L_i(a, \theta')$ for all $i \in N$, then $a \in F(\theta')$. Notice that strong monotonicity implies monotonicity, and monotonicity plus no veto power implies strong monotonicity. We have just shown that strong monotonicity is necessary for Nash implementation. In certain environments, it is also sufficient.

In the canonical mechanism of Section 3.2, if m^* is a "consensus" message profile such that rule 1 applies, i.e., all agents announce (a, θ) with $a \in F(\theta)$, then agent j's attainable set is $L_j(a, \theta)$. We have just seen why this may not work if no veto power is violated. The obvious solution is to modify rule 2 so that $C_j(a, \theta)$ becomes agent j's attainable set. If $n \geqslant 3$ and any linear ordering of A is an admissible preference relation ($\mathcal{P}_A \subseteq \mathcal{R}(\Theta)$) then this solution does work and strong monotonicity is sufficient for Nash implementation. A version of this result appears in Danilov (1992) [see also Moore (1992)]. It is instructive to prove it by comparing strong monotonicity to *condition M*, which is a necessary and (when $n \geqslant 3$) sufficient condition for Nash implementation in any environment [Sjöström (1991)][20]. The definition of condition M can be obtained from the definition of strong monotonicity by replacing the set $C_i(a, \theta)$ by a set $C_i^*(a, \theta)$ defined by Sjöström (1991). Since $C_i^*(a, \theta) \subseteq C_i(a, \theta)$ always holds, condition M implies strong monotonicity. But if $\mathcal{P}_A \subseteq \mathcal{R}(\Theta)$ and F is strongly monotonic, then $C_i^*(a, \theta) = C_i(a, \theta)$. Thus, if $\mathcal{P}_A \subseteq \mathcal{R}(\Theta)$ then strong monotonicity implies condition M, i.e., the two conditions are equivalent in this case.

There are two ways in which the definition of $C_i^*(a, \theta)$ differs from the definition of $C_i(a, \theta)$. The first difference is due to the fact that if F does not satisfy *unanimity*, then there are alternatives that must never be in the range of the outcome function h. Alternative a is a *problematic outcome* if and only if $a \notin F(\theta)$ for some state θ such that $L_i(a, \theta) = A$ for all $i \in N$. The problematic outcome a would clearly be a non-F-optimal Nash equilibrium outcome in state θ if $a = h(m)$ for some $m \in M$. After removing all problematic outcomes from A (several iterations may be necessary), what remains is some set $B^* \subseteq A$. Since we must have $h(m) \in B^*$ for all $m \in M$, Sjöström (1991) in effect treats B^* as the true "feasible set". His analogue of part (ii) of the definition of "awkward outcome" is therefore: for each $i \neq j$, $B^* \subseteq L_i(c, \theta')$. However, it turns out that this difference is irrelevant if $\mathcal{P}_A \subseteq \mathcal{R}(\Theta)$ [21].

[20] Condition M is equivalent to Moore and Repullo's (1990) condition μ. But it is easier to check.

[21] Suppose $\mathcal{P}_A \subseteq \mathcal{R}(\Theta)$ and let F be strongly monotonic. Let $a \in F(\theta)$, and let $\widehat{C}_j(a, \theta)$ be the set of outcomes in $L_j(a, \theta)$ that are not awkward according to the *new* definition (using B^* in (ii)). We

The second difference is due to the fact that, after removing the awkward outcomes from $L_j(a,\theta)$, we may discover a *second-order awkward outcome* $c \in C_j(a,\theta)$ such that for some $\theta' \in \Theta$: (i) $C_j(a,\theta) \subseteq L_j(c,\theta')$; (ii) for each $i \neq j$, $L_i(c,\theta') = A$; (iii) $c \notin F(\theta')$. Again, this would contradict implementation, so we must remove all second-order awkward outcomes from the attainable set, too. Indeed, Sjöström's (1991) algorithm may lead to iterated elimination of even higher-order awkward outcomes. When there are no more iterations to be made, what remains is the set $C_j^*(a,\theta) \subseteq C_j(a,\theta)$. It turns out that if $\mathcal{P}_A \subseteq \mathcal{R}(\Theta)$ and F is strongly monotonic, then there are no second-order awkward outcomes: the algorithm terminates after one step with $C_j^*(a,\theta) = C_j(a,\theta)$ [22]. In this case, strong monotonicity implies condition M, which is sufficient for Nash implementation [23]. Thus, if $n \geqslant 3$ and $\mathcal{P}_A \subseteq \mathcal{R}(\Theta)$ then the SCR F is Nash-implementable if and only if it is strongly monotonic, as claimed.

Consider two examples due to Maskin (1985). First, suppose $N = \{1,2,3\}$, $A = \{a,b,c\}$ and $\mathcal{R}(\Theta) = \mathcal{P}_A$. The SCR F is defined as follows. For any $\theta \in \Theta$, $a \in F(\theta)$ if and only if a majority prefers a to b, and $b \in F(\theta)$ if and only if a majority prefers b to a, and $c \in F(\theta)$ if and only if c is top-ranked in A by all agents. This SCR is monotonic and satisfies unanimity but not no veto power. Fix $j \in N$ and suppose θ is such that $bP_j(\theta)aP_j(\theta)c$, and $aP_i(\theta)b$ for all $i \neq j$. Then $F(\theta) = \{a\}$. Now suppose θ' is such that $bP_j(\theta')cP_j(\theta')a$ and $L_i(c,\theta') = A$ for all $i \neq j$. Since $L_j(a,\theta) = L_j(c,\theta') = \{a,c\}$ but $c \notin F(\theta')$, c is awkward in $L_j(a,\theta)$. Removing c, we obtain $C_j(a,\theta) = \{a\}$. By the symmetry of a and b, $C_j(b,\theta) = \{b\}$ whenever $aP_j(\theta)bP_j(\theta)c$ and $bP_i(\theta)a$ for all $i \neq j$. There are no other awkward outcomes and it can be verified that F is strongly monotonic, hence Nash-implementable. For a second example, consider any environment with $n \geqslant 3$, and let $a_0 \in A$ be a fixed "status quo" alternative. The *individually rational correspondence*, defined by $F(\theta) = \{a \in A: aR_i(\theta)a_0 \text{ for all } i \in N\}$, satisfies monotonicity and unanimity but not no veto power. If $a \in F(\theta)$ then $a_0 \in L_j(a,\theta)$ for all $j \in N$.

claim $\widehat{C}_j(a,\theta) = C_j(a,\theta)$. Clearly, $\widehat{C}_j(a,\theta) \subseteq C_j(a,\theta)$ since $B^* \subseteq A$. Thus, we only need to show $C_j(a,\theta) \subseteq \widehat{C}_j(a,\theta)$. Suppose $c \in L_j(a,\theta)$ but $c \notin \widehat{C}_j(a,\theta)$. Then there is θ' such that $L_j(a,\theta) \subseteq L_j(c,\theta')$ and $B^* \subseteq L_i(c,\theta')$ for each $i \neq j$, and $c \notin F(\theta')$. We claim $c \notin C_j(a,\theta)$. Suppose, in order to get a contradiction, that $c \in C_j(a,\theta)$. Then, if $\theta'' \in \Theta$ is a state where $L_j(a,\theta) = L_j(c,\theta'')$ and $L_i(c,\theta'') = A$ for each $i \neq j$, we have $c \in F(\theta'')$. It is easy to check that strong monotonicity implies $C_i(c,\theta'') \subseteq B^*$ for all i. Thus, $C_j(c,\theta'') \subseteq L_j(c,\theta'') \subseteq L_j(c,\theta')$ and $C_i(c,\theta'') \subseteq B^* \subseteq L_i(c,\theta')$ for each $i \neq j$, so $c \in F(\theta')$ by strong monotonicity. This is a contradiction. Thus, $C_j(a,\theta) \subseteq \widehat{C}_j(a,\theta)$.

[22] We claim that there are no second-order awkward outcomes if $\mathcal{P}_A \subseteq \mathcal{R}(\Theta)$ and F is strongly monotonic. Suppose $a \in F(\theta)$, $c \in C_j(a,\theta) \subseteq L_j(c,\theta')$, and for each $i \neq j$, $L_i(c,\theta') = A$. Since $\mathcal{P}_A \subseteq \mathcal{R}(\Theta)$ there exists $\theta'' \in \Theta$ such that $L_j(c,\theta'') = L_j(a,\theta)$ and $L_i(c,\theta'') = A$ for all $i \neq j$. Since $c \in C_j(a,\theta)$, we have $c \in F(\theta'')$. Now, $C_j(c,\theta'') = C_j(a,\theta) \subseteq L_j(c,\theta')$ and $L_i(c,\theta') = A$ for all $i \neq j$, so $c \in F(\theta')$ by strong monotonicity.

[23] Actually, since $C_i^*(a,\theta)$ is supposed to be agent i's attainable set at a Nash equilibrium m^* such that $h(m^*) = a \in F(\theta)$, Sjöström (1991) explicitly required $a \in C_i^*(a,\theta)$. Such a requirement is not explicit in strong monotonicity. But if $\mathcal{P}_A \subseteq \mathcal{R}(\Theta)$ and F is strongly monotonic then it is easy to check that $a \in C_i(a,\theta) = C_i^*(a,\theta)$.

If $c \in L_j(a, \theta) \subseteq L_j(c, \theta')$ and $L_i(c, \theta') = A$ for each $i \neq j$, then $cR_i(\theta')a_0$ for all $i \in N$ so $c \in F(\theta')$. Therefore, there are no awkward outcomes, and condition M and strong monotonicity both reduce to monotonicity. Since F is monotonic, it is Nash-implementable.

If $\mathcal{R}(\Theta) = \mathcal{R}_A$ then any monotonic F which satisfies Pareto indifference is strongly monotonic[24]. This fact is useful because if F is Nash-implementable when $\mathcal{R}(\Theta) = \mathcal{R}_A$ then implementation is possible (using the same mechanism) when the domain of preferences is restricted in an arbitrary way. In the context of voting, an even stronger symmetry condition called *neutrality* is often imposed. Neutrality requires that the SCR never discriminates among alternatives based on their labelling. Suppose $a \in F(\theta)$ and $c \in L_j(a, \theta)$, and state $\theta' \in \Theta$ is such that $L_j(a, \theta) \subseteq L_j(c, \theta')$ and for each $i \neq j$, $L_i(c, \theta') = A$. Let $\theta'' \in \Theta$ be a state where preferences are just as in θ' except for a permutation of alternatives a and c in the ranking of each agent[25]. Then $R_i(\theta'')$ is a monotonic transformation of $R_i(\theta)$ at a for each agent $i \in N$, so monotonicity would imply $a \in F(\theta'')$. The neutrality condition then requires that, in view of the symmetry of the two states θ' and θ'', $c \in F(\theta')$ so c is not awkward. But with no awkward outcomes monotonicity is equivalent to strong monotonicity. This yields a nice characterization of Nash-implementable neutral social choice rules.

Theorem 3: [Moulin (1983)]. *Suppose $n \geqslant 3$, and $\mathcal{R}(\Theta) = \mathcal{P}_A$ or $\mathcal{R}(\Theta) = \mathcal{R}_A$. Then a neutral SCR is Nash-implementable if and only if it is monotonic.*

Let $a \in F(\theta)$. Alternative $c \in L_i(a, \theta)$ is an *essential outcome for agent i in $L_i(a, \theta)$* if and only if there exists $\hat{\theta} \in \Theta$ such that $c \in F(\hat{\theta})$ and $L_i(c, \hat{\theta}) \subseteq L_i(a, \theta)$. Let $E_i(a, \theta) \subseteq L_i(a, \theta)$ denote the set of all outcomes that are essential for agent i in $L_i(a, \theta)$. An SCR F is *essentially monotonic* if and only if for all $(a, \theta, \theta') \in A \times \Theta \times \Theta$ the following is true: if $a \in F(\theta)$ and $E_i(a, \theta) \subseteq L_i(a, \theta')$ for all $i \in N$, then $a \in F(\theta')$. If F is monotonic then $E_i(a, \theta) \subseteq C_i(a, \theta)$ [26]. If $\mathcal{P}_A \subseteq \mathcal{R}(\Theta)$ then $C_i(a, \theta) \subseteq E_i(a, \theta)$ [27]. Thus, while essential monotonicity is in general stronger than strong monotonicity, the two conditions are equivalent if $\mathcal{P}_A \subseteq \mathcal{R}(\Theta)$.

Theorem 4: [Danilov (1992)]. *Suppose $n \geqslant 3$ and $\mathcal{P}_A \subseteq \mathcal{R}(\Theta)$. The SCR F is Nash-implementable if and only if it is essentially monotonic.*

[24] There are no awkward outcomes in this case. Indeed, let $a \in F(\theta)$, and suppose $c \in L_j(a, \theta) \subseteq L_j(c, \theta')$ and for each $i \neq j$, $L_i(c, \theta') = A$. We claim $c \in F(\theta')$. Let θ'' be such that for all $i \in N$, $cI_i(\theta'')a$ and for all $x, y \in A - \{c\}$, $xR_i(\theta'')y$ if and only if $xR_i(\theta)y$. Since $a \in F(\theta)$, monotonicity implies $a \in F(\theta'')$. Pareto indifference implies $c \in F(\theta'')$. But $L_i(c, \theta'') = L_i(a, \theta) \cup \{c\} \subseteq L_i(c, \theta')$ for all i, so $c \in F(\theta')$ by monotonicity.

[25] To make use of the neutrality condition we need to assume that the preference domain $\mathcal{R}(\Theta)$ is rich enough that such permutations are admissible. Of course, this is true if $\mathcal{R}(\Theta) = \mathcal{P}_A$ or $\mathcal{R}(\Theta) = \mathcal{R}_A$.

[26] If $c \in E_j(a, \theta)$ then there is $\hat{\theta} \in \Theta$ such that $c \in F(\hat{\theta})$ and $L_j(c, \hat{\theta}) \subseteq L_j(a, \theta)$. If $L_j(a, \theta) \subseteq L_j(c, \theta')$ and $L_i(c, \theta') = A$ for each $i \neq j$, then $c \in F(\theta')$ by monotonicity. Hence, $c \in C_j(a, \theta)$.

[27] If $c \in C_j(a, \theta)$ then $c \in F(\hat{\theta})$ for $\hat{\theta} \in \Theta$ such that $L_j(c, \hat{\theta}) = L_j(a, \theta)$ and $L_i(c, \hat{\theta}) = A$ for all $i \neq j$. So $c \in E_j(a, \theta)$.

Yamato (1992) has shown that essential monotonicity is a sufficient condition for Nash implementation in any environment (when $n \geqslant 3$), but it is a necessary condition only if $\mathcal{R}(\Theta)$ is sufficiently large.

3.4. Weak implementation

If $\widetilde{F}(\theta) \subseteq F(\theta)$ for all $\theta \in \Theta$ then $\widetilde{F}$ is a *subcorrespondence* of F, denoted $\widetilde{F} \subseteq F$. To weakly implement the SCR F is equivalent to fully implementing a non-empty valued subcorrespondence of F. Fix an SCR F, and for all $\theta \in \Theta$ define

$$F^*(\theta) \equiv \{a \in A: a \in F(\tilde{\theta}) \text{ for all } \tilde{\theta} \in \Theta \text{ such that } L_i(a, \theta) \subseteq L_i(a, \tilde{\theta}) \text{ for all } i \in N\}.$$

Theorem 5. *If $F^*(\theta) \neq \emptyset$ for all $\theta \in \Theta$ then F^* is a monotonic SCR.*

Proof: Suppose $a \in F^*(\theta)$ and $L_i(a, \theta) \subseteq L_i(a, \theta')$ for all $i \in N$. Suppose $\tilde{\theta} \in \Theta$ is such that $L_i(a, \theta') \subseteq L_i(a, \tilde{\theta})$ for all $i \in N$. Then $L_i(a, \theta) \subseteq L_i(a, \theta') \subseteq L_i(a, \tilde{\theta})$ for all i. Since $a \in F^*(\theta)$ we must have $a \in F(\tilde{\theta})$. Therefore, $a \in F^*(\theta')$. □

If $F^*(\theta) = \emptyset$ for some $\theta \in \Theta$ then F does not have any monotonic subcorrespondence, but if $F^*(\theta) \neq \emptyset$ for all $\theta \in \Theta$ then F^* is the maximal monotonic subcorrespondence of F. Moreover, F is monotonic if and only if $F^* = F$. Now, suppose that $n \geqslant 3$. If $F^*(\theta) \neq \emptyset$ for all $\theta \in \Theta$ and F satisfies no veto power, then F^* satisfies no veto power too and is Nash-implementable by Theorem 2, hence F is weakly implementable. Conversely, if F is weakly Nash-implementable, then Theorem 1 implies that F has a monotonic non-empty valued subcorrespondence $\widehat{F} \subseteq F$. Then $\widehat{F} \subseteq F^*$ so $F^*(\theta) \neq \emptyset$ for all $\theta \in \Theta$. Summarizing, we have the following.

Theorem 6. *If F can be weakly Nash-implemented then $F^*(\theta) \neq \emptyset$ for all $\theta \in \Theta$. Conversely, if $n \geqslant 3$ and F satisfies no veto power and $F^*(\theta) \neq \emptyset$ for all $\theta \in \Theta$, then F can be weakly Nash-implemented (and F^* is the maximal Nash-implementable subcorrespondence of F).*

3.5. Strategy-proofness and rich domains of preferences

We next show that there is an intimate connection between Nash-implementability and strategy-proofness of an SCF, when the preference domain has a "product structure" and is either "rich" or consists of strict orderings.

The preference domain $\mathcal{R}(\Theta)$ has a product structure if it takes the form $\mathcal{R}(\Theta) = \times_{i=1}^{n} \mathcal{R}_i$. For any coalition $C \subseteq N$ and any $R \in \mathcal{R}(\Theta)$, we write $R = (R_C, R_{-C})$ where $R_C \equiv \{R_i\}_{i \in C} \in \mathcal{R}_C(\Theta) \equiv \times_{i \in C} \mathcal{R}_i$ and $R_{-C} \in \times_{i \notin C} \mathcal{R}_i$. We also define $R_C(\theta) \equiv \{R_i(\theta)\}_{i \in C}$ and $R_{-C}(\theta) \equiv \{R_i(\theta)\}_{i \notin C}$ for any $\theta \in \Theta$. If the SCF f is ordinal, as it will be if it is monotonic, then the mapping $\bar{f}: \mathcal{R}(\Theta) \to A$ such that $\bar{f}(R(\theta)) = f(\theta)$ for all $\theta \in \Theta$ is well defined. An ordinal SCF f on a domain with a product structure is *strategy-proof* if, for all $i \in N$, all $\theta \in \Theta$, and all

$R_i' \in \mathcal{R}_i(\Theta), u_i(\bar{f}(R_i, R_{-i}), \theta) \geqslant u_i(\bar{f}(R_i', R_{-i}), \theta)$, where $(R_i, R_{-i}) = (R_i(\theta), R_{-i}(\theta))$. An ordinal SCF f on a domain with a product structure is *coalitionally strategy-proof* if, for all $\theta \in \Theta$, all non-empty coalitions $C \subseteq N$, and all preferences $R_C' \in \mathcal{R}_C(\Theta)$, there exists $i \in C$ such that

$$u_i\left(\bar{f}\left(R_C, R_{-C}\right), \theta\right) \geqslant u_i\left(\bar{f}\left(R_C', R_{-C}\right), \theta\right), \tag{1}$$

where $(R_C, R_{-C}) = (R_C(\theta), R_{-C}(\theta))$. Note that coalitional strategy-proofness implies ordinary strategy-proofness. If the SCF f is strategy-proof, then the revelation mechanism $\Gamma = \langle \times_{i=1}^n \mathcal{R}_i, \bar{f} \rangle$ has the property that, for any $i \in N$ and any $\theta \in \Theta$, truthfully reporting $R_i = R_i(\theta)$ is agent i's dominant strategy in state θ. If in addition f is coalitionally strategy-proof, then no coalitional deviation from truth-telling can make all members of the deviating coalition strictly better off.

To define "rich domain", we first introduce the concept of "improvement". If $u_i(a, \theta) \geqslant u_i(b, \theta)$ and $u_i(a, \theta') \leqslant u_i(b, \theta')$ and at least one inequality is strict, then b *improves with respect to a for agent i as the state changes from θ to θ'*. The following condition was introduced by Dasgupta, Hammond and Maskin (1979).

Definition. *Rich domain*: For any $a, b \in A$ and any $\theta, \theta' \in \Theta$, if, for all $i \in N$, b does not improve with respect to a for when the state changes from θ to θ', then there exists $\theta'' \in \Theta$ such that $L_i(a, \theta) \subseteq L_i(a, \theta'')$ and $L_i(b, \theta') \subseteq L_i(b, \theta'')$ for all $i \in N$.

Theorem 7: [Dasgupta, Hammond and Maskin (1979)]. *Suppose f is a monotonic SCF, the domain is rich, and the preference domain has a product structure $\mathcal{R}(\Theta) = \times_{i=1}^n \mathcal{R}_i$. Then f is coalitionally strategy-proof.*

Proof: Let f be as hypothesized. Let $C \subseteq N$ be any coalition. Suppose that the true preference profile in state θ is $R = (R_C, R_{-C}) = R(\theta)$. Consider a preference profile $R' = R(\theta') = (R_C', R_{-C})$, with $R_i' \neq R_i$ for $i \in C$ and $R_i' = R_i$ for $i \notin C$. Let $a = f(\theta) = \bar{f}(R_C, R_{-C})$ and $b = f(\theta') = \bar{f}(R_C', R_{-C})$. If $a = b$ then Inequality (1) holds trivially for all $i \in C$, so suppose $a \neq b$.

We claim that there exists $i \in C$ such that b improves with respect to a for agent i as the state changes from θ to θ'. Notice that because $R_i' = R_i$ for $i \notin C$, b cannot improve with respect to a for any such agent. Hence, if the claim is false, the definition of rich domain implies that there exists $\theta'' \in \Theta$ such that $L_i(a, \theta) \subseteq L_i(a, \theta'')$ and $L_i(b, \theta') \subseteq L_i(b, \theta'')$ for all $i \in N$. But then, from monotonicity, we have $a = f(\theta'')$ and $b = f(\theta'')$, a contradiction of f's single-valuedness. Hence the claim holds after all.

But b improving with respect to a for agent $i \in C$ implies that

$$u_i\left(\bar{f}\left(R_C, R_{-C}\right), \theta\right) \geqslant u_i\left(\bar{f}\left(R_C', R_{-C}\right), \theta\right),$$

and so f is coalitionally strategy-proof as claimed. □

Theorem 8: [Dasgupta, Hammond and Maskin (1979)]. *Suppose that $n \geqslant 3$. If $\mathcal{R}(\Theta)$ has a product structure and consists of strict orderings ($\mathcal{R}(\Theta) \subseteq \mathcal{P}_A$) and f is a strategy-proof SCF satisfying no veto power, then f is Nash-implementable.*

Proof: Let f be as hypothesized. We claim that f is monotonic. Suppose that, for some $\theta, \theta' \in \Theta$ and $a \in A$, we have $a = f(\theta)$ and $L_i(a, \theta) \subseteq L_i(a, \theta')$ for all $i \in N$. Let $R = R(\theta)$ and $R' = R(\theta')$. Because $\mathcal{R}(\Theta)$ has a product structure, there exists a state $\theta'' \in \Theta$ such that $(R'_1, R_2, \ldots, R_n) = R(\theta'')$. Let $c = f(\theta'')$. If $c \neq a$, then because f is strategy-proof, $u_1(a, \theta) > u_1(c, \theta)$ and $u_1(c, \theta') > u_1(a, \theta')$. But the former inequality implies that $c \in L_1(a, \theta)$ and, hence, from hypothesis, $c \in L_1(a, \theta')$, a contradiction of the latter inequality. Thus, $a = c$, after all. We conclude that $a \in f(\theta'')$, and, repeating the same argument for $i = 2, \ldots, n$, that $a \in f(\theta')$. Thus, f is indeed monotonic. Theorem 2 then implies that f is Nash-implementable. □

3.6. Unrestricted domain of strict preferences

Suppose society has to make a choice from a finite set A. The set of admissible preferences is the set of all linear orderings, $\mathcal{R}(\Theta) = \mathcal{P}_A$. This domain is rich, and so Theorem 7 applies. The SCR F is *dictatorial on its image* if and only if there exists $i \in N$ such that for all $\theta \in \Theta$ and all $a \in F(\theta)$, $u_i(a, \theta) \geqslant u_i(b, \theta)$ for all $b \in F(\Theta)$.

Theorem 9: [Gibbard (1973), Satterthwaite (1975)]. *Suppose that A is a finite set, $\mathcal{R}(\Theta) = \mathcal{P}_A$, and f is a strategy-proof SCF such that $f(\Theta)$ contains at least three alternatives. Then f is dictatorial on its image.*

Theorem 10: [Muller and Satterthwaite (1977), Dasgupta, Hammond and Maskin (1979), Roberts (1979)]. *Suppose the SCF f is Nash-implementable, A is a finite set, $f(\Theta)$ contains at least three alternatives, and $\mathcal{R}(\Theta) = \mathcal{P}_A$. Then f is dictatorial on its image.*

Proof: By Theorem 1 f is monotonic. By Theorem 7 f is strategy-proof. By Theorem 9 it must be dictatorial on its image. □

Theorem 10 is false without the hypothesis of single-valuedness. For example, the weak Pareto correspondence is monotonic and satisfies no veto power in any environment, so by Theorem 2, it can be Nash-implemented (when $n \geqslant 3$). Theorems 9 and 10 are also false without the hypothesis that the image contains at least three alternatives. To see this, let $N(a, b, \theta)$ denote the number of agents who strictly prefer a to b in state θ. Suppose $A = \{x, y\}$ and define the *method of majority rule* as follows: $F(\theta) = \{x\}$ if $N(x, y, \theta) > N(y, x, \theta)$, $F(\theta) = \{y\}$ if $N(x, y, \theta) < N(y, x, \theta)$, and $F(\theta) = \{x, y\}$ if $N(x, y, \theta) = N(y, x, \theta)$. If n is odd and $\mathcal{R}(\Theta) = \mathcal{P}_A$ then F is single-valued, monotonic, and satisfies no veto power. By Theorem 2 it can be Nash-implemented and by Theorem 7 it is coalitionally strategy-proof.

When A contains at least three alternatives the results are mainly negative. The plurality rule (which picks the alternative that is top-ranked by the greatest number of agents) is not monotonic, and neither are other well-known voting rules such as the Borda and Copeland rules. None of these social choice rules can be even

weakly Nash-implemented when $|A| \geqslant 3$. Peleg (1998) showed that all monotonic and strongly unanimous SCRs violate Sen's (1970) condition of *minimal liberty*. Indeed, if $\mathcal{R}(\Theta) = \mathcal{P}_A$ then monotonicity and strong unanimity imply Pareto optimality[28], but Sen showed that no Pareto optimal SCR can satisfy minimal liberty.

3.7. Economic environments

An interesting environment is the *L-good exchange economy* $\langle A_E, N, \Theta_E \rangle$. In this environment no veto power is automatically satisfied when $n \geqslant 3$, since $n-1$ non-satiated agents can never agree on the best way to allocate the social endowment. Thus, monotonicity will be both necessary and sufficient for implementation when $n \geqslant 3$. The feasible set is

$$A_E = \left\{ a = (a_1, a_2, \ldots, a_n) \in \mathbb{R}^L_+ \times \mathbb{R}^L_+ \times \cdots \times \mathbb{R}^L_+ : \sum_{i=1}^{n} a_i \leqslant \omega \right\},$$

where $a_i \in \mathbb{R}^L_+$ is agent i's consumption vector, and $\omega \in \mathbb{R}^L_{++}$ is the aggregate endowment vector[29]. Let $A^0_E = \{a \in A_E : a_i \neq 0 \text{ for all } i \in N\}$ denote the set of allocations where no agent gets a zero consumption vector. Each agent cares only about his own consumption and strictly prefers more to less. Although preferences are defined only over *feasible* allocations in A_E, it is conventional to introduce utility functions defined on $\mathbb{R}^L_+$. Thus, in each state $\theta \in \Theta_E$, for each agent $i \in N$ there is a continuous, increasing[30] and strictly quasi-concave function $v_i(\cdot, \theta)$: $\mathbb{R}^L_+ \to \mathbb{R}$ such that $u_i(a, \theta) = v_i(a_i, \theta)$ for all $a \in A$. Moreover, for any function from $\mathbb{R}^L_+$ to $\mathbb{R}$ satisfying these standard assumptions, there is a state $\theta \in \Theta_E$ such that agent i's preferences are represented by that function. The domain $\mathcal{R}_E \equiv \mathcal{R}(\Theta_E)$, which consists of all preference profiles that can be represented by utility functions satisfying these standard assumptions, is rich [Dasgupta, Hammond and Maskin (1979)]. By Theorem 7, monotonicity implies strategy-proofness for single-valued social choice rules. If $n = 2$ then strategy-proofness plus Pareto optimality implies dictatorship in this environment [Zhou (1991)][31]. Strategy-proof, Pareto optimal and non-dictatorial social

[28] For suppose $u_i(a, \theta) > u_i(b, \theta)$ for all $i \in N$ but $b \in F(\theta)$. Consider the state θ' where preferences are as in state θ except that a has been moved to the top of everybody's preference. Then, $R_i(\theta')$ is a monotonic transformation of $R_i(\theta)$ at b for all i so $b \in F(\theta')$ by monotonicity, but $F(\theta') = \{a\}$ by strong unanimity, a contradiction.

[29] $\mathbb{R}^L$ is L-dimensional Euclidean space, $\mathbb{R}^L_+ = \{x \in \mathbb{R}^L : x_k \geqslant 0, \text{ for } k = 1, \ldots, L\}$ and $\mathbb{R}^L_{++} = \{x \in \mathbb{R}^L : x_k > 0, \text{ for } k = 1, \ldots, L\}$.

[30] A function $v_i(\cdot, \theta)$ is increasing if and only if $v_i(a_i, \theta) > v_i(a'_i, \theta)$ whenever $a_i \geqslant a'_i$, $a_i \neq a'_i$.

[31] Of course, these results depend on the assumptions we make about admissible preferences. Suppose $n = L = 2$ and let $\Theta^* \subset \Theta_E$ be such that in each state $\theta \in \Theta^*$ both goods are normal for both agents. Let ℓ be a fixed "downward sloping line" that passes through the Edgeworth box. For each $\theta \in \Theta^*$ let $f(\theta)$ be the unique Pareto optimal and feasible point on ℓ. Then f: $\Theta^* \to A_E$ is a monotonic, Pareto optimal and non-dictatorial SCF which (using the mechanism described in Section 3.8) can be Nash implemented in the environment $\langle A_E, \{1, 2\}, \Theta^* \rangle$.

choice functions exist when $n \geqslant 3$, but they are not very attractive [Satterthwaite and Sonnenschein (1981)]. More positive results are obtained if the requirement of single-valuedness is relaxed. Hurwicz (1979a) and Schmeidler (1980) constructed simple "market mechanisms" where each agent proposes a consumption vector and a price vector, and the set of Nash equilibrium outcomes coincides with the set of Walrasian outcomes. Reichelstein and Reiter (1988) showed (under certain smoothness conditions on the outcome function) that the minimal dimension of the message space M of any such mechanism is approximately $n(L-1)+L/(n-1)$ [32]. However, the mechanisms in these articles violated the feasibility constraint $h(m) \in A$ for all $m \in M$. In fact, the Walrasian correspondence W is not monotonic, hence not Nash-implementable, in the environment $\langle A_E, N, \Theta_E \rangle$. The problem occurs because a change in preferences over *non-feasible* consumption bundles can eliminate a Walrasian equilibrium on the boundary of the feasible set. The minimal monotonic extension of the Walrasian correspondence W is the *constrained Walrasian correspondence* W^c [Hurwicz, Maskin and Postlewaite (1995)]. For simple, feasible and continuous implementation of the constrained Walrasian correspondence, see Postlewaite and Wettstein (1989) and Hong (1995). Under certain assumptions, any Nash-implementable SCR must contain W^c as a sub-correspondence [Hurwicz (1979b), Thomson (1979)].

Hurwicz (1960, 1972) discussed "proposed outcome" mechanisms where each agent i's message m_i is his proposed net trade vector. "Information smuggling" can be ruled out by requiring that in equilibrium $h(m) = m$. In exchange economies, a proposed trade vector does not in general contain enough information about marginal rates of substitution to ensure a Pareto efficient outcome [Saijo, Tatamitani and Yamato (1996) and Sjöström (1996a)], although the situation may be rather different in production economies with known production sets [Yoshihara (2000)]. Dutta, Sen and Vohra (1995) characterized the class of SCRs that can be implemented by "elementary" mechanisms where agents propose prices as well as trade vectors. This class contains the Walrasian correspondence (on their preference domain, $W = W^c$).

For public goods economies, Hurwicz (1979a) and Walker (1981) constructed simple mechanisms such that the set of Nash equilibrium outcomes coincides with the set of Lindahl outcomes. Again, however, $h(m) \notin A$ was allowed out of equilibrium. In Walker's mechanism each agent announces a real number for each of the K public goods, so the dimension of M is nK, the minimal dimension of any smooth Pareto efficient mechanism in this environment [Sato (1981), Reichelstein and Reiter (1988)]. Like the Walrasian correspondence, the Lindahl correspondence is not monotonic in general. The minimal monotonic extension is the *constrained Lindahl correspondence*, nicely implemented by Tian (1989).

[32] The first term $n(L-1)$ is due to each agent proposing an $(L-1)$-dimensional consumption vector for himself, and the second term $L/(n-1)$ comes from the need to also allow announcements of price variables. Smoothness conditions are needed to rule out "information smuggling" [Hurwicz (1972), Mount and Reiter (1974), Reichelstein and Reiter (1988)].

In many economic environments a *single crossing condition* holds which makes monotonicity rather easy to satisfy. For example, suppose there is a seller and a buyer, a divisible good and "money". Let q denote the transfer of money from the buyer to the seller (which can be positive or negative), and $x \geqslant 0$ the amount of the good delivered from the seller to the buyer. The feasible set is $A = \{(q,x) \in \mathbb{R}^2 : x \geqslant 0\}$. The state of the world is denoted $\theta = (\theta_s, \theta_b) \in [0,1] \times [0,1] \equiv \Theta$. The seller's payoff function is $u(q,x,\theta_s)$, with $\partial u/\partial q > 0$, $\partial u/\partial x < 0$. The buyer's payoff function is $v(q,x,\theta_b)$, with $\partial v/\partial q < 0$, $\partial v/\partial x > 0$. An increase in θ_s represents an increase in the seller's marginal production cost, and an increase in θ_b represents an increase in the buyer's marginal valuation. More formally, the single crossing condition states that

$$\frac{\partial}{\partial \theta_s}\left|\frac{\partial u/\partial x}{\partial u/\partial q}\right| > 0 \quad \text{and} \quad \frac{\partial}{\partial \theta_b}\left|\frac{\partial v/\partial x}{\partial v/\partial q}\right| > 0.$$

Under this assumption, a monotonic transformation can only take place at a boundary allocation where $x = 0$. Monotonicity says that if $(q,0) \in F(\theta_s,\theta_b)$, $\theta_s' \geqslant \theta_s$ and $\theta_b' \leqslant \theta_b$, then $(q,0) \in F(\theta_s',\theta_b')$.

3.8. Two agent implementation

The necessary and sufficient condition for two-agent Nash implementation in general environments was given by Moore and Repullo (1990) and Dutta and Sen (1991b). To see why the case $n = 2$ may be more difficult than the case $n \geqslant 3$ note that rule 2 of the canonical mechanism for Nash implementation singles out a unique deviator from a "consensus". However, with $n = 2$ this is not possible. Let $a \in F(\theta)$ and $a' \in F(\theta')$. If Γ Nash-implements F then there are message profiles $(m_1,m_2) \in N^\Gamma(\theta)$ and $(m_1',m_2') \in N^\Gamma(\theta')$ such that $h(m_1,m_2) = a$ and $h(m_1',m_2') = a'$. Since agent 1 should have no incentive to deviate to message m_1 in state θ' and agent 2 should have no incentive to deviate to message m_2' in state θ, a property called *weak non-empty lower intersection* must be satisfied: there exists an outcome $b = h(m_1,m_2')$ such that $a'R_1(\theta')b$ and $aR_2(\theta)b$. In most economic environments this condition automatically holds, so the case $n = 2$ is similar to the case $n \geqslant 3$. In the two-agent exchange economy $\langle A_E, \{1,2\}, \Theta_E\rangle$ (defined in Section 3.7) an SCR F can be Nash-implemented if and only if it is monotonic and satisfies a very weak boundary condition [Sjöström (1991)]. In particular, suppose F is monotonic and never recommends a zero consumption vector to any agent. That is, $F(\Theta_E) \subseteq A_E^0$. It is easy to check that the following mechanism Nash-implements F. Each agent $i \in \{1,2\}$ announces an outcome $a^i = (a_1^i, a_2^i) \in A_E^0$, where a_j^i is a proposed consumption vector for agent j, and a state $\theta^i \in \Theta_E$. Thus, $m_i = (a^i, \theta^i) \in M_i \equiv A_E^0 \times \Theta_E$. Let $h_i(m)$ denote agent i's consumption vector. Set $h_i(m) = a_i^i$ if

$$m_1 = m_2 \quad \text{and} \quad a^i \in F(\theta^i),$$

or if

$$R_j(\theta^i) = R_j(\theta^j), \quad R_i(\theta^j) \neq R_i(\theta^i) \quad \text{and} \quad a^j R_i(\theta^j) a^i.$$

Otherwise, set $h_i(m) = 0$.

Such positive results for the case $n = 2$ do rely on restrictions on the domain of preferences, as the following result shows.

Theorem 11: [Maskin (1999), Hurwicz and Schmeidler (1978)]. *Suppose* $n = 2$ *and* $\mathcal{P}_A \subseteq \mathcal{R}(\Theta)$. *If the SCR* F *is weakly Pareto optimal and Nash-implementable, then* F *is dictatorial.*

Proof: Suppose a weakly Pareto optimal SCR F is implemented by $\Gamma = \langle M, h \rangle$. For any $a \in A$, there is an agent $i = i(a) \in \{1, 2\}$ such that a is always in his attainable set, i.e., $a \in h(m_j, M_i)$ for all $m_j \in M_j$ $(j \neq i)$. For if not, then there is $m \in M$ such that when m is played neither agent 1 nor agent 2 can attain a, but then $x = h(m)$ is a Pareto dominated Nash equilibrium outcome whenever both agents rank a first and x second. In fact, for any two outcomes a and b we must have $i(a) = i(b)$, for otherwise there is no Nash equilibrium when agent $i(a)$ ranks a first and agent $i(b)$ ranks b first. So there exists a dictator, i.e., an agent i such that $h(m_j, M_i) = A$ for all $m_j \in M_j$. □

4. Implementation with complete information: further topics

4.1. *Refinements of Nash equilibrium*

Message $m_i \in M_i$ is a *dominated strategy* in state $\theta \in \Theta$ for agent $i \in N$ if and only if there exists $m_i' \in M_i$ such that $u_i(h(m_{-i}, m_i'), \theta) \geqslant u_i(h(m_{-i}, m_i), \theta)$ for all $m_{-i} \in M_{-i}$, and $u_i(h(m_{-i}, m_i'), \theta) > u_i(h(m_{-i}, m_i), \theta)$ for some $m_{-i} \in M_{-i}$. A Nash equilibrium is an *undominated Nash equilibrium* if and only if no player uses a dominated strategy[33]. Notice that we are considering domination in the *weak* sense. It turns out that "almost anything" can be implemented in undominated Nash equilibria. Of course, a mechanism that implements a non-monotonic SCR F in undominated Nash equilibria must have non-F-optimal Nash equilibria involving dominated strategies. The assumption here, however, is that dominated strategies will never be used.

An SCR F satisfies *property* Q if and only if, for all $(\theta, \theta') \in \Theta \times \Theta$ such that $F(\theta) \not\subseteq F(\theta')$, there exists an agent $i \in N$ and two alternatives $(a, b) \in A \times A$ such that b improves with respect to a for agent i as the state changes from θ to θ', and moreover this agent i is not indifferent over all alternatives in A in state θ'. Property Q is a very weak condition because it only involves a preference reversal over two arbitrary alternatives a and b, neither of which has to be F-optimal. If no agent is ever indifferent over all alternatives in A, then property Q is equivalent to ordinality.

[33] The Nash equilibria of the the canonical mechanism for Nash implementation are not necessarily undominated, because if $a \in F(\theta)$ is the worst outcome in A for agent i in state θ then it may be a (weakly) dominated strategy for him to announce a. However, Yamato (1999) modified the canonical mechanism so that all Nash equilibria are undominated. He showed that if $n \geqslant 3$ then any Nash implementable SCR is doubly implementable in Nash and undominated Nash equilibria.

Theorem 12: [Palfrey and Srivastava (1991)]. *If the SCR F is implementable in undominated Nash equilibria, then it satisfies property Q. Conversely, if $n \geqslant 3$ and F satisfies property Q and no veto power, then F is implementable in undominated Nash equilibria.*

Proof: It is not difficult to see the necessity of property Q. To prove the sufficiency part, we will simplify by assuming that (i) $\mathcal{R}(\Theta)$ has a product structure, $\mathcal{R}(\Theta) = \times_{i=1}^{n} \mathcal{R}_i$, and (ii) *value distinction* holds: for all $i \in N$ and all ordered pairs $(R_i, R_i') \in \mathcal{R}_i \times \mathcal{R}_i$, if $R_i' \neq R_i$ then there exist outcomes b and c in A such that cR_ib and $bP_i'c$. Let F satisfy property Q and no veto power. Then F is ordinal, so we can suppose it is defined directly on the set of possible preference profiles, F: $\mathcal{R} \equiv \times_{i=1}^{n} \mathcal{R}_i \rightarrow A$. Consider the following mechanism. Agent i's message space is

$$M_i = A \times \mathcal{R} \times \mathcal{R}_i \times Z \times Z \times Z,$$

where Z is the set of all positive integers. A typical message for agent i is $m_i = (a^i, R^i, r^i, z^i, \zeta^i, \gamma^i) \in M_i$, where $a^i \in A$ is an outcome, $R^i = (R_1^i, R_2^i, \ldots, R_n^i) \in \mathcal{R}$ is a statement about the preference profile, $r^i \in \mathcal{R}_i$ is another report about agent i's *own* preference, and (z^i, ζ^i, γ^i) are three integers. The outcome function is as follows[34].

Rule 1: If there exists $j \in N$ such that $(a^j, R^i) = (a, R)$ for all $i \neq j$, and $a \in F(R)$, then $h(m) = a$.

Rule 2: If rule 1 does not apply then:

(a) if there is $j \in N$ such that $j = (\sum_{k=1}^{n} z^k) \bmod(2n)$, set

$$h(m) = a^j;$$

(b) if there is $j \in N$ such that $n + j = (\sum_{k=1}^{n} z^k) \bmod(2n)$ and $\gamma^j > \zeta^{j-1}$, set

$$h(m) = \begin{cases} a^{j-1} & \text{if } a^{j-1} r^j a^{j+1} \\ a^{j+1} & \text{otherwise} \end{cases};$$

(c) if there is $j \in N$ such that $n + j = (\sum_{k=1}^{n} z^k) \bmod(2n)$ and $\gamma^j \leqslant \zeta^{j-1}$, set

$$h(m) = \begin{cases} a^{j-1} & \text{if } a^{j-1} R_j^j a^{j+1} \\ a^{j+1} & \text{otherwise} \end{cases}.$$

Notice that rule 1 includes the case of a consensus, $(a^i, R^i) = (a, R)$ for *all* i, as well as the case where a single agent j differs from the rest. Rule 2a is a modulo game similar to rule 3 of the canonical mechanism for Nash implementation. Rule 2b chooses

[34] References to agents $j-1$ and $j+1$ are always "modulo n" (if $j = 1$ then agent $j-1$ is agent n; if $j = n$ then agent $j+1$ is agent 1).

agent j's most preferred outcome among a^{j-1} and a^{j+1} according to preferences r^j, and rule 2c chooses agent j's most preferred outcome among a^{j-1} and a^{j+1} according to preferences R_j^j.

Let $R^* = (R_1^*, \ldots, R_n^*)$ denote the true preference profile. Let $U^\Gamma(R^*)$ denote the set of undominated Nash equilibria when the preference profile is R^*. The proof consists of several steps.

Step 1. If m_j is undominated for agent j then $r^j = R_j^*$. Indeed, r^j only appears in rule 2b, where "truthfully" announcing $r^j = R_j^*$ is always at least as good as any false announcement. By value distinction there exists a^{j-1} and a^{j+1} such that the preference is strict.

Step 2. If m_j is undominated for agent j then $R_j^j = R_j^*$. For, if $R_j^j \neq R_j^*$ then (since $r^j = R_j^*$ by step 1) if $n+j = (\sum_{k=1}^n z^k) \bmod(2n)$, agent j always weakly prefers rule 2b to rule 2c, and by value distinction there exists a^{j-1} and a^{j+1} such that this preference is strict. But increasing γ^j increases the chance of rule 2b at the expense of rule 2c, without any other consequence, so m_j cannot be undominated.

Step 3. If m is a Nash equilibrium then either $(a^i, R^i) = (a, R)$ for all $i \in N$ and $a \in F(R)$, or there is j such that for all $i \neq j$, $h(m) R_i^* a$ for all $a \in A$. This follows from rule 2a (the same argument was used in the canonical mechanism for Nash implementation).

Step 4. $h(U^\Gamma(R^*)) \subseteq F(R^*)$. For, if $m \in U^\Gamma(R^*)$, then by steps 1 and 2, $R_j^j = r^j = R_j^*$ for all j. By step 3, either rule 1 applies, in which case $(a^i, R^i) = (a, R^*)$ for all $i \in N$ and $h(m) = a \in F(R^*)$, or else $h(m) \in F(R^*)$ by no veto power.

Step 5. $F(R^*) \subseteq h(U^\Gamma(R^*))$. Each agent j announcing $(R^j, r^j) = (R^*, R_j^*)$ "truthfully" and $a^j = a \in F(R^*)$ (and three arbitrary integers) is an undominated Nash equilibrium. (Notice that if $R_j^j = r^j$ then there is no possibility that γ^j can change the outcome).

Steps 4 and 5 imply $h(U^\Gamma(R^*)) = F(R^*)$. □

A similar possibility result was obtained for *trembling-hand perfect Nash equilibria* by Sjöström (1991). If agents have strict preferences over an underlying finite set of basic alternatives B, and $A = \Delta(B)$ as discussed in Section 4.2, then a sufficient condition for F to be implementable in trembling-hand perfect equilibria is that F satisfies no veto power as well as its "converse": if all but one agent agree on which alternative is the worst, then this alternative is not F-optimal. Yamato (1993) considers *double* implementation in Nash and undominated Nash equilibrium.

A mechanism is *bounded* if and only if each dominated strategy is dominated by some *undominated* strategy [Jackson (1992)]. The mechanism used by Sjöström (1991) for trembling hand perfect Nash implementation has a finite message space, hence it is bounded. But Palfrey and Srivastava's (1991) mechanism for undominated Nash implementation contains infinite sequences of strategies dominating each other, hence it is not bounded. This is illustrated by step 2 of the proof of Theorem 12. However, in economic environments satisfying standard assumptions, any ordinal SCF which

never recommends a zero consumption vector to any agent can be implemented in undominated Nash equilibria by a very simple bounded mechanism which does not use integer or modulo games.

Theorem 13: [Jackson, Palfrey and Srivastava (1994), Sjöström (1994)]. *Consider the economic environment* $\langle A_E, N, \Theta_E \rangle$ *with* $n \geqslant 2$. *If* f *is an ordinal SCF such that* $f(\Theta_E) \subseteq A_E^0$ *then* f *can be implemented in undominated Nash equilibria by a bounded mechanism.*

Proof: We prove this for $n = 2$ using a mechanism due to Jackson, Palfrey and Srivastava (1994)[35]. If f is ordinal then without loss of generality we may assume f is defined on $\mathcal{R}_E$ instead of on Θ_E. Thus, consider $f: \mathcal{R}_E \to A_E^0$. Let $f_j(R)$ denote agent j's f-optimal consumption vector when the preference profile is R. Each agent $i \in \{1,2\}$ announces *either* a preference profile $R^i = (R_1^i, R_2^i) \in \mathcal{R}_E$, *or* a pair of outcomes $(a^i, b^i) \in A_E^0 \times A_E^0$. Notice that $a^i = (a_1^i, a_2^i)$ is a pair of consumption vectors, and $b^i = (b_1^i, b_2^i)$ is another pair. Let $h_j(m)$ denote agent j's consumption.

Rule 1: Suppose both agents announce a preference profile. If $R_j^i \neq R_j^j$, then $h_i(m) = 0$. If $R_j^i = R_j^j$, then $h_i(m) = f_i(R^j)$.

Rule 2: Suppose agent i announces a preference profile R^i and agent j announces outcomes (a^j, b^j). Then, $h_j(m) = 0$. If $a^j P_i^i b^j$ then $h_i(m) = a_i^j$, otherwise $h_i(m) = b_i^j$.

Rule 3: In all other cases, $h_1(m) = h_2(m) = 0$.

Suppose the true preference profile is $R^* = (R_1^*, R_2^*)$. It is a dominated strategy to announce outcomes, since that guarantees a zero consumption bundle. Moreover, truthfully announcing $R_i^i = R_i^*$ dominates lying since the only effect lying about his own preferences can have on agent i's consumption is to give him an inferior allocation under rule 2 [36]. Now, if agent j is announcing preferences, any best response for agent i must involve $R_j^i = R_j^j$. (Since utility functions are increasing, getting $f_i(R^j) \neq 0$ is strictly better than getting no consumption at all). Therefore, in the unique undominated Nash equilibrium both agents announce the true preference profile, so this mechanism implements f. □

The most disturbing feature of the mechanism in the proof of Theorem 13 is that agent i's only reason to announce $R_i^i = R_i^*$ truthfully is that it will give him a preferred outcome in case agent $j \neq i$ uses the dominated strategy of announcing outcomes. This problem does not occur in Sjöström's (1994) mechanism. In that mechanism, each agent reports a preference ordering for himself and two "neighbors", and the only dominated strategies are those where an agent does not tell the truth about himself.

[35] Sjöström's (1994) mechanism is similar but works only for $n \geqslant 3$.

[36] The allocation can be strictly inferior because value distinction holds in this environment. Indeed, since preferences are defined over *feasible* outcomes, if $R_i \neq R_i^*$ then there is $(a^j, b^j) \in A_E^0 \times A_E^0$ such that $a^j P_i^* b^j$ but $b^j R_i a^j$.

When these dominated strategies have been removed, a second round of elimination of *strictly* dominated strategies leads each agent to match what his neighbors are saying about themselves.

The *iterated* removal of dominated strategies was considered by Farquharson (1969) and Moulin (1979) in their analyses of dominance solvable voting schemes. Abreu and Matsushima (1994) showed that if the feasible set consists of lotteries over a set of basic alternatives, strict value distinction holds, and the social planner can use "small fines", then any SCF can be implemented using the iterated elimination of dominated strategies (without using integer and modulo games). It does not matter in which order dominated strategies are eliminated, but many rounds of elimination may be required [see Glazer and Rosenthal (1992) and Abreu and Matsushima (1992b)].

A Nash equilibrium is *strong* if and only if no group $S \subseteq N$ has a *joint* deviation which makes all agents in S better off. Monotonicity is a necessary condition for implementation in strong Nash equilibria [Maskin (1979b, 1985)]. A necessary and sufficient condition for strong Nash implementation was found by Dutta and Sen (1991a), and an algorithm for checking it was provided by Suh (1995). Moulin and Peleg (1982) established the close connection between strong Nash implementation and the notion of effectivity function. For *double* implementation in Nash and strong Nash equilibria, see Maskin (1979a, 1985), Schmeidler (1980) and Suh (1997). In the environment $\langle A_E, N, \Theta_E \rangle$ with $n \geqslant 2$, any monotonic and Pareto optimal SCR F such that $F(\Theta_E) \subseteq A_E^0$ can be doubly implemented in Nash and strong Nash equilibria, even if joint deviations may involve ex post trade of goods "outside the mechanism" [Maskin (1979a), Sjöström (1996b)]. Further results on implementation with coalition formation are contained in Peleg (1984) and Suh (1996).

4.2. *Virtual implementation*

Virtual implementation was first studied by Abreu and Sen (1991) and Matsushima (1988). Let B be a finite set of "basic alternatives", and let the set of feasible outcomes be $A = \Delta(B)$, the set of all probability distributions over B. The elements of $\Delta(B)$ are called *lotteries*. Let $\Delta^0(B)$ denote the subset of $\Delta(B)$ which consists of all lotteries that give strictly positive probability to all alternatives in B. Let $d(a, b)$ denote the Euclidean distance between lotteries $a, b \in \Delta(B)$. Two SCRs F and G are *ε-close* if and only if for all $\theta \in \Theta$ there exists a bijection α_θ: $F(\theta) \to G(\theta)$ such that $d(a, \alpha_\theta(a)) \leqslant \varepsilon$ for all $a \in F(\theta)$. An SCR F is *virtually Nash-implementable* if and only if for all $\varepsilon > 0$ there exists an SCR G which is Nash-implementable and ε-close to F. If F is virtually implemented, then the social planner accepts a strictly positive probability that the equilibrium outcome is some undesirable element of B. However, this probability can be made arbitrarily small.

Theorem 14: [Abreu and Sen (1991), Matsushima (1988)]. *Suppose $n \geqslant 3$. Let B be a finite set of "basic alternatives" and let the set of feasible alternatives be $A = \Delta(B)$. Suppose for all $\theta \in \Theta$, no agent is indifferent over all alternatives in*

B, and preferences over A satisfy the von Neumann–Morgenstern axioms. Then any ordinal SCR F: $\Theta \to A$ is virtually Nash-implementable.

Proof: Since any ordinal SCR F: $\Theta \to \Delta(B)$ can be approximated arbitrarily closely by an ordinal SCR G such that $G(\Theta) \subseteq \Delta^0(B)$, it suffices to show that any such G is Nash-implementable. So let G: $\Theta \to \Delta^0(B)$ be an ordinal SCR. In the environment $\langle \Delta^0(B), N, \Theta \rangle$ the SCR G satisfies no veto power because no agent has a most preferred outcome in $\Delta^0(B)$. If $a \in G(\theta)$ but $a \notin G(\theta')$, then since G is ordinal there is $i \in N$ such that $R_i(\theta) \neq R_i(\theta')$. The von Neumann–Morgenstern axioms imply that indifference surfaces are hyperplanes, so $R_i(\theta')$ cannot be a monotonic transformation of $R_i(\theta)$ at $a \in \Delta^0(B)$. Thus, G is monotonic. By Theorem 2, G is Nash-implementable in environment $\langle \Delta^0(B), N, \Theta \rangle$. But then G is also Nash-implementable when the feasible set is $\Delta(B)$, since we can always just disregard the alternatives that are not in $\Delta^0(B)$. □

Of course, if an SCR is not ordinal then it cannot be virtually Nash-implemented, so ordinality is both necessary and sufficient under the hypothesis of Theorem 14[37]. The proof of Theorem 14 does not do justice to the work of Abreu and Sen (1991) and Matsushima (1988), since their mechanisms are better behaved than the canonical mechanism. For virtual implementation using iterated elimination of strictly dominated strategies, see Abreu and Matsushima (1992a).

4.3. Mixed strategies

A mixed strategy μ_i for agent $i \in N$ is a probability distribution over M_i. For simplicity, we restrict attention to mixed strategies that put positive probability on only a finite number of messages. Let $\mu_i(m_i)$ denote the probability that agent i sends message m_i, let $\mu(m) \equiv \times_{i=1}^{n} \mu_i(m_i)$ and $\mu_{-j}(m_{-j}) \equiv \times_{i \neq j} \mu_i(m_i)$. In most of the implementation literature, only the pure strategy equilibria of the mechanism are verified to be F-optimal, leaving open the possibility that there may be non-F-optimal mixed strategy equilibria[38]. In particular, in the proof of Theorem 2 we did not establish that all mixed strategy Nash equilibria are F-optimal. In fact they need not be. To see the problem, consider a mixed strategy Nash equilibrium $\mu = (\mu_1, \ldots, \mu_n)$ for the canonical mechanism in state θ^*. Suppose $\mu(m) > 0$ for m such that rule 2 applies, that is,

$$(a^i, \theta^i) = (a, \theta) \quad \text{for all } i \neq j, \tag{2}$$

but $(a^j, \theta^j) \neq (a, \theta)$. If $\mu(m) = 1$ then $h(m)$ must be top-ranked by each agent $i \neq j$. Otherwise, agent $i \neq j$ could induce his favorite alternative $\hat{a}^i$ via rule 3. Thus, no

[37] Recall that ordinality says that only preferences over A matter for the social choice. Here, $A = \Delta(B)$.

[38] Exceptions include Abreu and Matsushima (1992a), Jackson, Palfrey and Srivastava (1994) and Sjöström (1994).

veto power guarantees $h(m) \in F(\theta^*)$. But suppose $\mu_{-i}(m'_{-i}) > 0$ for some m'_{-i} such that $m'_k = (a', \theta', z'_k)$ for all $k \neq i$, where $a' \in F(\theta')$ and

$$u_i\left(\hat{a}^i, \theta'\right) > u_i\left(a', \theta'\right) > u_i\left(a, \theta'\right). \tag{3}$$

Then, although agent i can induce $\hat{a}^i$ when the others play m_{-i}, Inequality (3) and rule 2 of the canonical mechanism imply that he cannot induce $\hat{a}^i$ when the others play m'_{-i}. Indeed, if he tries to do so the outcome will be a', which in state θ^* may be much worse for him than a (the outcome that, from Inequality (3) and rule 2, he would get by sticking to m_i). Hence, he may prefer not to try to induce $\hat{a}^i$ even if he strictly prefers it to $h(m)$. And so we cannot infer that $h(m)$ is F-optimal. The difficulty arises because which message is best for agent i to send depends on the messages that the other agents send, but if the other agents are using mixed strategies then agent i is unable to forecast (except probabilistically) what these messages will be. Nevertheless, the canonical mechanism can be readily modified to take account of mixed strategies.

Suppose $n \geqslant 3$. The following is a version of a modified canonical mechanism proposed by Maskin (1999). A typical message for agent i is $m_i = (a^i, \theta^i, z^i, \alpha^i)$, where $a^i \in A$ is an outcome, $\theta^i \in \Theta$ is a state, $z^i \in Z$ is a positive integer, and $\alpha^i: A \times \Theta \to A$ is a mapping from outcomes and states to outcomes satisfying $\alpha^i(a, \theta) \in L_i(a, \theta)$ for all (a, θ). Let the outcome function be defined as follows.

Rule 1: Suppose there exists $j \in N$ such that $(a^i, \theta^i, z^i) = (a, \theta, 1)$ for all $i \neq j$ and $z^j = 1$. Then $h(m) = a$.

Rule 2: Suppose there exists $j \in N$ such that $(a^i, \theta^i, z^i) = (a, \theta, 1)$ for all $i \neq j$ and $z^j > 1$. Then $h(m) = \alpha^j(a, \theta)$.

Rule 3: In all other cases let $h(m) = a^i$ for i such that $z^i \geqslant z^j$ for all $j \in N$ (if there are several such i, choose the one with the lowest index i).

Notice that rule 1 encompasses the case of a consensus, $(a^i, \theta^i, z^i) = (a, \theta, 1)$ for all $i \in N$. The mapping α^i enables agent i, in effect, to propose a *contingent* outcome, which eliminates the difficulty noted above. Indeed, for any mixed Nash equilibrium μ, agent i has nothing to lose from setting $\alpha^i(a, \theta)$ equal to his favorite outcome in $L_i(a, \theta)$, a^i equal to his favorite outcome in all of A, and z^i larger than any integer announced with positive probability by any other agent[39]. Such a strategy guarantees that he gets his favorite outcome in his attainable set $L_i(a, \theta)$ whenever $(a^k, \theta^k, z^k) = (a, \theta, 1)$ for all $k \neq i$, and for all other m_{-i} such that $\mu_{-i}(m_{-i}) > 0$ it will cause him to win the integer game in rule 3. Thus, in Nash equilibrium, if $\mu(m) > 0$ and rule 1 applies to m, so $(a^i, \theta^i) = (a, \theta)$ for all i, then $h(m) = a$ must be the most preferred alternative in $L_i(a, \theta)$ for each agent i. But if instead rule 2 or rule 3 applies to m then $h(m)$ must be top-ranked in all of A by at least $n - 1$ agents. Thus, if F

[39] If such favorite outcomes do not exist, the argument is more roundabout but still goes through. The same is true if the other agents use mixed strategies with infinite support. In that case, agent i cannot guarantee that he will have the highest integer, but he can make the probability arbitrarily close to one and that is all we need.

is monotonic and satisfies no veto power then $\mu(m) > 0$ implies $h(m)$ is F-optimal. Conversely, if $a \in F(\theta)$ then there is a pure strategy Nash equilibrium in state θ where $(a^i, \theta^i, z^i) = (a, \theta, 1)$ for all $i \in N$ [40]. So this mechanism Nash-implements F even when we take account of mixed strategies.

Maskin and Moore (1999) show that the extensive form mechanisms considered by Moore and Repullo (1988) and Abreu and Sen (1990) can also be suitably modified for mixed strategies. We conjecture that analogous modifications can be made for mechanisms corresponding to most of the other solution concepts that have been considered in the literature.

4.4. Extensive form mechanisms

An SCR F is *implementable in subgame-perfect equilibria* if and only if there exists an extensive form mechanism such that in each state $\theta \in \Theta$, the set of subgame-perfect equilibrium outcomes equals $F(\theta)$. Extensive form mechanisms were studied by Farquharson (1969) and Moulin (1979). Moore and Repullo (1988) obtained a partial characterization of subgame-perfect implementable SCRs. Their result was improved on by Abreu and Sen (1990).

To illustrate the ideas that are involved, consider a *quasi-linear* environment with two agents, $N = \{1, 2\}$. There is an underlying set B of "basic alternatives", which can be finite or infinite. In addition, a good called "money" can be used to freely transfer utility between the agents. Let y_i denote the net transfer of money to agent i, which can be positive or negative. However, we assume social choice rules are *bounded*: they do not recommend arbitrarily large transfers to or from any agent. A typical outcome is denoted $a = (b, y_1, y_2)$. The feasible set is

$$A = \{(b, y_1, y_2) \in B \times \mathbb{R} \times \mathbb{R}: y_1 + y_2 \leqslant 0\}.$$

Notice that $y_1 + y_2 < 0$ is allowed (money can be destroyed or given to some outside party). In all states, each agent i's payoff function is of the quasi-linear form $u_i(a, \theta) = v_i(b, \theta) + y_i$, where v_i is bounded. Assume *strict value distinction* in the sense that we can select $(b(\theta, \theta'), y(\theta, \theta')) \in B \times \mathbb{R}$, for each ordered pair $(\theta, \theta') \in \Theta \times \Theta$, such that the following is true. Whenever $\theta \neq \theta'$, there exists a "test agent" $j = j(\theta, \theta') = j(\theta', \theta) \in N$ that experiences a strict preference reversal of the form:

$$v_j\left(b\left(\theta, \theta'\right), \theta\right) + y\left(\theta, \theta'\right) > v_j\left(b\left(\theta', \theta\right), \theta\right) + y\left(\theta', \theta\right), \tag{4}$$

and

$$v_j\left(b\left(\theta, \theta'\right), \theta'\right) + y\left(\theta, \theta'\right) < v_j\left(b\left(\theta', \theta\right), \theta'\right) + y\left(\theta', \theta\right). \tag{5}$$

In this environment, any bounded SCF $f: \Theta \to A$ can be implemented in subgame-perfect equilibria by the following simple two-stage mechanism. [See Moore and

[40] The Nash equilibrium strategies are undominated as long as a is neither the best nor the worst outcome in A for any agent.

Repullo (1988) and Moore (1992) for similar mechanisms.] Stage 1 consists of simultaneous announcements of a state: each agent $i \in N$ announces $\theta^i \in \Theta$. If $\theta^1 = \theta^2 = \theta$ then the game ends with the outcome $f(\theta)$. If $\theta^1 \neq \theta^2$, then go to stage 2. Let $j(1) = j(\theta^1, \theta^2)$ denote the "test agent" for (θ^1, θ^2), let $\theta = \theta^{j(1)}$ denote the test agent's announcement in stage 1 and let $\theta' = \theta^{j(0)}$ denote the announcement made by the other agent, agent $j(0) \neq j(1)$. Let $a(1) = (b(\theta, \theta'), y_1, y_2)$ with $y_{j(1)} = y(\theta, \theta') - z$ and $y_{j(0)} = -z$ where $z > 0$. Let $a(2) = (b(\theta', \theta), y_1, y_2)$ with $y_{j(1)} = y(\theta', \theta) - z$ and $y_{j(0)} = r > 0$. In stage 2, agent $j(1)$ decides the outcome of the game by choosing either $a(1)$ or $a(2)$. By Inequalities (4) and (5), agent $j(1)$ prefers $a(2)$ to $a(1)$ if θ' is the true state, but he prefers $a(1)$ to $a(2)$ if θ is the true state. In effect, agent $j(0)$'s announcement θ' is "confirmed" if agent $j(1)$ chooses $a(2)$, and then agent $j(0)$ receives a "bonus" r. But if agent $j(1)$ chooses $a(1)$, then agent $j(0)$ pays a "fine" z. Agent $j(1)$ pays the fine whichever outcome he chooses in stage 2 (this does not affect his preference reversal over $a(1)$ and $a(2)$).

If the agents disagree in stage 1, then at least one agent must pay the fine z. This is incompatible with equilibrium if z is sufficiently big, because any agent can avoid the fine by agreeing with the other agent in stage 1[41]. Thus, in equilibrium both agents will announce the same state, say $\theta^1 = \theta^2 = \theta$, in stage 1. Suppose the *true* state is $\theta' \neq \theta$. Let $j(1) = j(\theta, \theta')$ be the test agent for (θ, θ'). Suppose agent $j(0) \neq j(1)$ deviates in stage 1 by announcing $\theta^{j(0)} = \theta'$ truthfully. In stage 2, agent $j(1)$ will choose $a(2)$ so agent $j(0)$ will get the bonus r which makes him strictly better off if r is sufficiently big. Thus, if z and r are big enough, in any subgame-perfect equilibrium both agents must announce the *true* state in stage 1. Conversely, both agents announcing the true state in stage 1 is part of a subgame-perfect equilibrium which yields the f-optimal outcome (no agent wants to deviate, because he will pay the fine if he does). Thus, f is implemented in subgame-perfect equilibria. The reader can verify that the sequences $a(0) = f(\theta), a(1), a(2)$ in A, and $j(0), j(1)$ in N, fulfil the requirements of the following definition (with $\ell = 1$ and $A' = A$).

Definition. *Property* α: There exists a set A', with $F(\Theta) \subseteq A' \subseteq A$, such that for all $(a, \theta, \theta') \in A \times \Theta \times \Theta$ the following is true. If $a \in F(\theta) - F(\theta')$ then there exists a sequence of outcomes $a(0) = a, a(1), \ldots, a(\ell), a(\ell+1)$ in A' and a sequence of agents $j(0), j(1), \ldots, j(\ell)$ in N such that:

(i) for $k = 0, 1, \ldots, \ell$,

$$u_{j(k)}(a(k), \theta) \geqslant u_{j(k)}(a(k+1), \theta);$$

(ii)

$$u_{j(\ell)}(a(\ell), \theta') < u_{j(\ell)}(a(\ell+1), \theta');$$

[41] As long as f and v_i are bounded, each agent prefers any $f(\theta)$ to paying a large fine. Without boundedness, z and r would have to depend on (θ, θ').

(iii) for $k = 0, 1, \ldots, \ell$, in state θ' outcome $a(k)$ is not the top-ranked outcome in A' for agent $j(k)$;
(iv) if in state θ', $a(\ell + 1)$ is the top-ranked outcome in A' for each agent $i \neq j(\ell)$, then either $\ell = 0$ or $j(\ell - 1) \neq j(\ell)$.

If F is monotonic then $a \in F(\theta) - F(\theta')$ implies the existence of $(a(1), j(0)) \in A \times N$ such that $u_{j(0)}(a, \theta) \geqslant u_{j(0)}(a(1), \theta)$ and $u_{j(0)}(a, \theta') < u_{j(0)}(a(1), \theta')$, so sequences satisfying (i)–(iv) exist (with $\ell = 0$). Hence, property α is weaker than monotonicity. Recall that property Q requires that someone's preferences reverse over two arbitrary alternatives. Since property α requires a preference reversal over two alternatives $a(\ell)$ and $a(\ell + 1)$ that can be connected to a by sequences satisfying (i)–(iv), property α is stronger than property Q.

Theorem 15: [Moore and Repullo (1988), Abreu and Sen (1990)]. *If the SCR F is implementable in subgame-perfect equilibria, then it satisfies property α. Conversely, if $n \geqslant 3$ and F satisfies property α and no veto power, then F is implementable in subgame-perfect equilibria.*

Recently, Vartiainen (1999) found a condition which is both necessary and sufficient for subgame-perfect implementation when $n \geqslant 3$ and A is a finite set. Herrero and Srivastava (1992) derived a necessary and sufficient condition for an SCF to be implementable via backward induction using a finite game of perfect information. An interesting connection between extensive and normal form implementation is drawn by Glazer and Rubinstein (1996).

4.5. Renegotiation

So far we have been assuming implicitly that the mechanism Γ is immutable. In this section we shall allow for the possibility that agents might *renegotiate* it. Articles on implementation theory are often written as though an exogenous planner simply imposes the mechanism on the agents. But this is not the only possible interpretation of the implementation setting. The agents might choose the mechanism *themselves*, in which case we can think of the mechanism as a "constitution", or a "contract" that the agents have signed. Suppose that when this contract is executed (i.e., when the mechanism is played) it results in a Pareto inefficient outcome. Presumably, if the contract has been properly designed, this could not occur in equilibrium: agents would not deliberately design an inefficient contract. But inefficient outcomes might be incorporated in contracts as "punishments" for *deviations* from equilibrium. However, if a deviation from equilibrium has occurred, why should the agents accept the corresponding outcome given that it is inefficient? Why can't they "tear up" their contract (abandon the mechanism) and sign a new one resulting in a Pareto superior outcome? In other words, why can't they *renegotiate*? But if punishment is renegotiated, it may no longer serve as an effective deterrent to deviation from equilibrium. Notice that renegotiation would normally not pose a problem if all that

mattered was that the final outcome should be Pareto optimal. However, a contract will in general try to achieve a particular *distribution* of the payoffs (for example, in order to share risks), and there is no reason why renegotiation would lead to the desired distribution. Thus, the original contract must be designed with the possibility of renegotiation explicitly taken into account. Our discussion follows Maskin and Moore (1999). A different approach is suggested by Rubinstein and Wolinsky (1992).

Consider the following example, drawn from Maskin and Moore (1999). Let $N = \{1, 2\}$, $\Theta = \{\theta, \theta'\}$, and $A = \{a, b, c\}$. Agent 1 always prefers a to c to b. Agent 2 has preferences $cP_2(\theta)aP_2(\theta)b$ in state θ and $bP_2(\theta')aP_2(\theta')c$ in state θ'. Suppose $f(\theta) = a$ and $f(\theta') = b$. If we leave aside the issue of renegotiation for the moment, there is a simple mechanism that Nash-implements this SCF, namely, agent 2 chooses between a and b. He prefers a in state θ and b in state θ' and so f will be implemented. But what if he happened to choose b in state θ? Since b is Pareto dominated by a and c the agents will be motivated to renegotiate. If, in fact, b were renegotiated to a, there would be no problem since whether agent 2 chose a or b in state θ, the final outcome would be $a = f(\theta)$. However, if b were renegotiated to c in state θ, then agent 2 would intentionally choose b in state θ, anticipating the renegotiation to c. Then b would not serve to punish agent 2 for deviating from the choice he is supposed to make in state θ, and the simple mechanism would no longer work. Moreover, from Theorem 16 below, no other mechanism can implement f either. Thus, renegotiation can indeed constrain the SCRs that are implementable. But the example also makes clear that whether or not f is implementable depends on the precise nature of renegotiation (if b is renegotiated to a, implementation is possible; if b is renegotiated to c, it is not). Thus, rather than speaking merely of the "implementation of f", we should speak of the "implementation of f for a given renegotiation process".

In this section the feasible set is $A = \Delta(B)$, the set of all probability distributions over a set of basic alternatives B. We identify degenerate probability distributions that assign probability one to some basic alternative b with the alternative b itself. The renegotiation process can be expressed as a function r: $B \times \Theta \to B$, where $r(b, \theta)$ is the (basic) alternative to which the agents renegotiate in state $\theta \in \Theta$ if the fall-back outcome (i.e., the outcome prescribed by the mechanism) is $b \in B$. Assume renegotiation is *efficient* (for all b and θ, $r(b, \theta)$ is Pareto optimal in state θ) and *individually rational* (for all b and θ, $r(b, \theta)R_i(\theta)b$ for all i)[42]. For each $\theta \in \Theta$, define a function r_θ: $B \to B$ by $r_\theta(b) \equiv r(\theta, b)$. Let $x \in A$, assume for the moment that B is a finite set, and let $x(b)$ denote the probability that the lottery x assigns to outcome $b \in B$. Extend r_θ to lotteries in the following way: let $r_\theta(x) \in A$ be the lottery which assigns probability $\sum x(a)$ to basic alternative $b \in B$, where the sum is over

[42] Jackson and Palfrey (2001) propose an alternative set of assumptions. If in state θ any agent can *veto* the outcome of the mechanism and instead enforce an alternative $a(\theta)$, renegotiation will satisfy $r(b, \theta) = b$ if $bR_i(\theta)a(\theta)$ for all $i \in N$, and $r(b, \theta) = a(\theta)$ otherwise. In an exchange economy, $a(\theta)$ may be the endowment point, in which case the constrained Walrasian correspondence is not implementable [Jackson and Palfrey (2001)].

the set $\{a\colon r_\theta(a) = b\}$. For B an infinite set, define $r_\theta(x)$ in the obvious analogous way. Thus, we now have $r_\theta\colon A \to A$ for all $\theta \in \Theta$. Finally, given a mechanism $\Gamma = \langle M, h\rangle$ and a state $\theta \in \Theta$, let $r_\theta \circ h$ denote the composition of r_θ and h. That is, for any $m \in M$, $(r_\theta \circ h)(m) \equiv r_\theta(h(m))$. The composition $r_\theta \circ h\colon M \to A$ describes the *de facto* outcome function in state θ, since any basic outcome prescribed by the mechanism will be renegotiated according to r_θ. Notice that if the outcome $h(m)$ is a non-degenerate randomization over B, then renegotiation takes place *after* the uncertainty inherent in $h(m)$ has been resolved and the mechanism has prescribed a basic alternative in B. Let $\mathcal{S}(\langle M, r_\theta \circ h\rangle, \theta)$ denote the set of $\mathcal{S}$-equilibrium outcomes in state θ, when the outcome function h has been replaced by $r_\theta \circ h$. A mechanism $\Gamma = \langle M, h\rangle$ is said to *$\mathcal{S}$-implement the SCR F for renegotiation function r* if and only if $\mathcal{S}(\langle M, r_\theta \circ h\rangle, \theta) = F(\theta)$ for all $\theta \in \Theta$. In this section we restrict our attention to social choice rules that are *essentially single-valued*: for all $\theta \in \Theta$, if $a \in F(\theta)$ then $F(\theta) = \{b \in A\colon bI_i(\theta)a \text{ for all } i \in N\}$.

Much of implementation theory with renegotiation has been developed for its application to bilateral contracts. With $n = 2$, a simple set of conditions are necessary for implementation *regardless* of the refinement of Nash equilibrium that is adopted as the solution concept.

Theorem 16: [Maskin and Moore (1999)]. *The two-agent SCR F can be implemented in Nash equilibria (or any refinement thereof) for renegotiation function r only if there exists a random function $\tilde{a}\colon \Theta \times \Theta \to A$ such that,*

(i) for all $\theta \in \Theta$,

$$r_\theta(\tilde{a}(\theta, \theta)) \in F(\theta);$$

(ii) and for all $(\theta, \theta') \in \Theta \times \Theta$,

$$r_\theta(\tilde{a}(\theta, \theta))\, R_1(\theta)\, r_\theta\left(\tilde{a}\left(\theta', \theta\right)\right);$$

(iii) and

$$r_\theta(\tilde{a}(\theta, \theta))\, R_2(\theta)\, r_\theta\left(\tilde{a}\left(\theta, \theta'\right)\right).$$

If $\tilde{a}(\theta, \theta)$ is the (random) equilibrium outcome of a mechanism in state θ, then condition (i) ensures that the renegotiated outcome is F-optimal, and conditions (ii) and (iii) ensure that neither agent 1 nor agent 2 will wish to deviate and act as though the state were θ'.

The reason for introducing randomizations over basic alternatives in Theorem 16 and the following results is to enhance the possibility of punishing agents for deviating from equilibrium. By assumption, agents will always renegotiate to a Pareto optimal alternative. Thus, if agent 1 is to be punished for a deviation (i.e., if his utility is to be reduced below the equilibrium level), then agent 2 must, in effect, be rewarded for

this deviation (i.e., his utility must be raised above the equilibrium), once renegotiation is taken into account. But as we noted in Section 3.8, determining which agent has deviated may not be possible when $n = 2$, so it may be necessary to punish *both* agents. However, this cannot be done if one agent is always rewarded when the other is punished. That is where randomization comes in. Although, for each realization $b \in B$ of the random variable $\tilde{a} \in A$, $r_\theta(b)$ is Pareto optimal, the random variable $r_\theta(\tilde{a})$ need not be Pareto optimal (if the Pareto frontier in utility space is not linear). Hence, deliberately introducing randomization is a way to create mutual punishments despite the constraint of renegotiation.

In the case of a linear Pareto frontier[43] randomization does not help. In that case, the conditions of Theorem 16 become *sufficient* for implementation.

Theorem 17: [Maskin and Moore (1999)]. *Suppose that the Pareto frontier is linear for all $\theta \in \Theta$. Then the two-agent F can be implemented in Nash equilibria for renegotiation function r if there exists a random function $\tilde{a}$: $\Theta \times \Theta \to A$ satisfying conditions (i), (ii) and (iii) of Theorem 16.*

Under the hypothesis of Theorem 17, a mechanism in effect induces a two-person zero-sum game (renegotiation ensures that outcomes are Pareto efficient, and the linearity of the Pareto frontier means that payoffs sum to a constant). In zero-sum games, any refined Nash equilibrium must yield both players the same payoffs as all other Nash equilibria. Theorems 16 and 17 show that using refinements will not be helpful for implementation in such a situation.

With "quasi-linear preferences" the Pareto frontier is linear, and Segal and Whinston (2002) have shown that Theorem 17 can be re-expressed in terms of first-order conditions[44].

Theorem 18: [Segal and Whinston (2002)]. *Assume*

(i) $N = \{1, 2\}$;

(ii) the set of alternatives is

$$A = \{(b, y_1, y_2) \in B \times \mathbb{R} \times \mathbb{R}: y_1 + y_2 = 0\},$$

where B is a connected compact space;

(iii) $\Theta = [\underline{\theta}, \overline{\theta}]$ *is a compact interval in* $\mathbb{R}$; *and*

(iv) in each state $\theta \in \Theta$, each agent i's post-renegotiation preferences take the form: for all $(b, y_1, y_2) \in A$,

$$u_i(r_\theta(b, y_1, y_2), \theta) = v_i(b, \theta) + y_i,$$

where v_i is C^1.

[43] Formally, the frontier is linear in state θ if, for all $b, b' \in B$ that are both Pareto optimal in state θ, the lottery $\lambda b + (1 - \lambda) b'$ is also Pareto optimal, where λ is the probability of b.

[44] Notice that their feasible set is different from what we otherwise assume in this section.

If the SCR $F: \Theta \to A$ is implementable in Nash equilibria (or any refinement thereof) for renegotiation function r, then there exists $\hat{b}: \Theta \to B$ such that, for all $\theta \in \Theta$ and all $i \in N$,

$$u_i(F(\theta), \theta) = \int_{\underline{\theta}}^{\theta} \frac{\partial v_i}{\partial \theta}\left(\hat{b}(t), t\right) \mathrm{d}t + u_i\left(F\left(\underline{\theta}\right), \underline{\theta}\right). \tag{6}$$

Furthermore, if there is $i \in N$ such that $(\partial^2 v_i / \partial\theta\partial b)(b, \theta) > 0$ for all $b \in B$ and all $\theta \in \Theta$, then the existence of $\hat{b}$ satisfying Inequality (6) is sufficient for F's Nash-implementability by a mechanism where only agent i sends a message.

Notice that as F is essentially single-valued, we may abuse notation and write $u_i(F(\theta), \theta)$ in Inequality (6).

When the Pareto frontier is not linear it becomes possible to punish both agents for deviations from equilibrium. We obtain the following result for implementation in subgame-perfect equilibria.

Theorem 19: [Maskin and Moore (1999)]. *The two-agent SCR F can be implemented in subgame-perfect equilibria with renegotiation function r if there exists a random function $\tilde{a}: \Theta \to A$ such that*

(i) for all $\theta \in \Theta$, $r(\tilde{a}(\theta), \theta) \in F(\theta)$;

(ii) for all $(\theta, \theta') \in \Theta \times \Theta$ such that $r(\tilde{a}(\theta), \theta') \notin F(\theta')$ there exists an agent k and a pair of random alternatives $\tilde{b}(\theta, \theta')$, $\tilde{c}(\theta, \theta')$ in A such that

$$r\left(\tilde{b}\left(\theta, \theta'\right), \theta\right) R_k(\theta) r\left(\tilde{c}\left(\theta, \theta'\right), \theta\right),$$

and

$$r\left(\tilde{c}\left(\theta, \theta'\right), \theta'\right) P_k\left(\theta'\right) r\left(\tilde{b}\left(\theta, \theta'\right), \theta'\right);$$

(iii) if $Z \subseteq A$ is the union of all $\tilde{a}(\theta)$ for $\theta \in \Theta$ together with all $\tilde{b}(\theta, \theta')$ and $\tilde{c}(\theta, \theta')$ for $\theta, \theta' \in \Theta$, then no alternative $z \in Z$ is maximal for any agent i in any state $\theta \in \Theta$ even after renegotiation (that is, there exists some $\mathrm{d}^i(\theta) \in A$ such that $\mathrm{d}^i(\theta) P_i(\theta) r(z, \theta)$); and

(iv) there exists some random alternative $\tilde{e} \in A$ such that, for any agent i and any state $\theta \in \Theta$, every alternative in Z is strictly preferred to $\tilde{e}$ after renegotiation (that is, $r(z, \theta) P_i(\theta) r(\tilde{e}, \theta)$ for all $z \in Z$).

The definition of implementation with renegotiation suggests that characterization results should be r-translations of those for implementation when renegotiation is ruled out. That is, for each result *without* renegotiation, we can apply r to obtain the corresponding result *with* renegotiation. This is particularly clear if Nash equilibrium is the solution concept. From Theorems 1 and 2 we know that monotonicity is the key to Nash implementation. By analogy, we would expect that some form of "renegotiation-monotonicity" should be the key when renegotiation is admitted. More precisely, we

say that the SCR F is *renegotiation monotonic for renegotiation function* r provided that, for all $\theta \in \Theta$ and all $x \in F(\theta)$ there is $a \in A$ such that $r(a, \theta) = x$, and if $L_i(r(a, \theta), \theta) \subseteq L_i(r(a, \theta'), \theta')$ for all $i \in N$ then $r(a, \theta') \in F(\theta')$.

Theorem 20: [Maskin and Moore (1999)]. *The SCR F can be implemented in Nash equilibria with renegotiation function r only if F satisfies renegotiation monotonicity for r. Conversely, if $n \geqslant 3$ and no alternative is maximal in A for two or more agents, then F is implementable in Nash equilibria with renegotiation function r if F satisfies renegotiation monotonicity for r.*

By analogy with Section 4.1, Nash equilibrium refinements should allow the implementation of social choice rules that do not satisfy renegotiation monotonicity. Theorem 16 has in fact put substantial limits on what can be achieved when $n = 2$. But the situation when $n \geqslant 3$ is very different, at least in economic environments. Introducing a third party into a bilateral economic relationship makes it possible to simultaneously punish both original parties by transferring resources to the third party, which makes the problem of renegotiation much less serious[45]. Before stating this result formally, we need a definition. A renegotiation function r: $A_E \times \Theta_E \to A_E$ satisfies *disagreement point monotonicity* if for all $i \in N$, all $\theta \in \Theta_E$ and all $a, b \in A_E$ such that all agents except i get no consumption ($a_j = b_j = 0$ for all $j \neq i$), it holds that $r(a, \theta) R_i(\theta) r(b, \theta)$ if and only if $aR_i(\theta)b$. That is, if two fall-back outcomes a and b both give zero consumption to everyone except agent i, then agent i prefers to renegotiate from whichever fall-back outcome gives him higher utility. Standard bargaining solutions such as the Nash solution and the Kalai–Smorodinsky solution satisfy this property.

Theorem 21: [Sjöström (1999)]. *Consider the environment $\langle A_E, N, \Theta_E \rangle$ with $n \geqslant 3$. Let r be any renegotiation function that satisfies disagreement point monotonicity and individual rationality. If f is an ordinal and Pareto optimal SCF such that $f(\Theta_E) \subseteq A_E^0$, then f can be implemented in undominated Nash equilibria with renegotiation function r.*

Sjöström's (1999) mechanism is "non-parametric" in the sense that it does not depend on r. Moreover, it is both bounded and robust to collusion. It is sometimes argued that introducing a third party into a bilateral relationship may lead to collusion between the third party and one of the original parties. However, all undominated Nash equilibria of Sjöström's (1999) mechanism are coalition-proof, which is the appropriate solution concept when agents can collude but cannot write binding side-contracts ex ante (allowing binding ex ante agreements would take the analysis into the realm of n-person cooperative game theory). A possibility result similar to Sjöström's (1999)

[45] What is important is not that the third person knows the true state of the world, only that he is willing to accept transfers of goods from the original parties.

was obtained by Baliga and Brusco (2000) for implementation using extensive form mechanisms.

4.6. The planner as a player

The canonical mechanism for Nash implementation can be given the following intuitive explanation. Rule 1 states that if (a, θ) is a consensus among the agents, where $a \in F(\theta)$, then the outcome is a. Rule 2 states that agent j's attainable set at the consensus is the lower contour set $L_j(a, \theta)$. By "objecting" against the consensus, agent j can induce any $a^j \in L_j(a, \theta)$. Monotonicity is the condition that makes such objections effective. For if $\theta' \neq \theta$ is the true state and $a \notin F(\theta')$, then by monotonicity some agent j strictly prefers to deviate from the consensus with an objection $a^j \in L_j(a, \theta) - L_j(a, \theta')$. Agent j would have no reason to propose a^j in state θ since $a^j \in L_j(a, \theta)$, but he does have such an incentive in state θ' since $a^j \notin L_j(a, \theta')$.

Now suppose the mechanism is controlled by a social planner who does not know the true state of the world. She gets payoff $u_0(a, \theta)$ from alternative a in state θ, and the SCR F she wants to implement is

$$F(\theta) \equiv \arg\max_{a \in A} u_0(a, \theta). \tag{7}$$

Suppose F is Nash-implementable and the planner uses the canonical mechanism to implement it. By Inequality (7), the equilibrium outcome maximizes her payoff in each state of the world. But out of equilibrium, she faces a credibility problem similar to the one discussed in the previous section. After hearing out of equilibrium messages, she may want to change the rules that she herself has laid down. Specifically, consider the "objection" made by agent j which was described in the previous paragraph. Let $\Theta' = \{\theta' \in \Theta: a^j \notin L_j(a, \theta')\}$ be the set of states where agent j strictly prefers a^j to a. If player j tries to induce a^j via rule 2, when all the other agents are announcing (a, θ), then [following the logic of Farrell (1993) and Grossman and Perry (1986)] the planner's beliefs about the true state should be some probability distribution over Θ'. But a^j may not maximize the planner's expected payoff for any such beliefs, in which case she prefers to "tear up" the mechanism after agent j has made his objection. In this sense the outcome function may not be credible. The situation is even worse if the "modulo game" in rule 3 is triggered. Rule 3 may lead to zero consumption for everybody except the winner of the modulo game, but that may be an outcome the planner dislikes regardless of her beliefs about the state. If the planner cannot commit to carrying out "incredible threats" such as giving no consumption to $n - 1$ agents, then the implementation problem is very difficult. Conditions under which the planner can credibly implement the SCR given by Inequality (7) are discussed by Chakravorty, Corchón and Wilkie (1997) and Baliga, Corchón and Sjöström (1997).

On the other hand, if the planner can commit to the outcome function then explicitly allowing her to participate as a player in the game *expands* the set of implementable social choice rules. Consider a utilitarian social planner with payoff function

$$u_0(a, \theta) = \sum_{i=1}^{n} u_i(a, \theta).$$

The SCR F she wants to implement is the utilitarian SCR which is not even ordinal (it is not invariant to multiplying an agent's utility function by a scalar). If the planner does not play then this F cannot be implemented using any non-cooperative solution concept (even virtually). However, suppose the environment is $\langle A_E, N, \Theta_E \rangle$ with $n \geqslant 3$. If we let the planner, who does not know the true θ, participate in the mechanism by sending a message of her own, then the utilitarian SCR can be implemented in Bayesian Nash equilibria for "generic" prior beliefs over Θ [Baliga and Sjöström (1999)]. This does not quite contradict the fact that only ordinal social choice rules can be implemented. Inequality (7) implies that if $F(\theta) \neq F(\theta')$ then the planner's preferences over A must differ in states θ and θ', so *all* social choice rules are ordinal if the planner's own preferences are taken into account[46].

5. Bayesian implementation

Now we drop the assumption that each agent knows the true state of the world and consider the case of *incomplete information*.

5.1. Definitions

A generic state of the world is denoted $\theta = (\theta_1, \ldots, \theta_n)$, where θ_i is agent i's *type*. Let Θ_i denote the finite set of possible types for agent i, and $\Theta \equiv \Theta_1 \times \cdots \times \Theta_n$. Agent i knows his own type θ_i but may be unsure about $\theta_{-i} \equiv (\theta_1, \ldots, \theta_{i-1}, \theta_{i+1}, \ldots, \theta_n)$. Agent i's payoff depends only on his own type and the final outcome (*private values*). Thus, if the outcome is $a \in A$ and the state of the world is $\theta = (\theta_1, \ldots, \theta_n) \in \Theta$, then we will write agent i's payoff as $u_i(a, \theta_i)$ rather than $u_i(a, \theta)$. There exists a common prior distribution on Θ, denoted p. Conditional on knowing his own type θ_i, agent i's posterior distribution over $\Theta_{-i} \equiv \times_{j \neq i} \Theta_j$ is denoted $p(\cdot \mid \theta_i)$. It can be deduced from p using Bayes' rule for any θ_i which occurs with positive probability. If g: $\Theta_{-i} \to A$

[46] Hurwicz (1979b) considered the possibility of using an "auctioneer" whose payoff function agrees with the SCR. However, he considered Nash equilibria among the $n + 1$ players, which implicitly requires the auctioneer to know the true θ (or else relies on some adjustment process as discussed in the Introduction).

is any function, and $\theta_i \in \Theta_i$, then the expectation of $u_i(g(\theta_{-i}), \theta_i)$ conditional on θ_i is denoted

$$E\{u_i(g(\theta_{-i}), \theta_i) \mid \theta_i\} = \sum_{\theta_{-i} \in \Theta_{-i}} p(\theta_{-i} \mid \theta_i)\, u_i(g(\theta_{-i}), \theta_i).$$

A strategy profile in the mechanism $\Gamma = \langle M, h\rangle$ is denoted $\sigma = (\sigma_1, \ldots, \sigma_n)$, where for each i, σ_i: $\Theta_i \rightarrow M_i$ is a function which specifies the messages sent by agent i's different types. The message profile sent at state θ is denoted $\sigma(\theta) = (\sigma_1(\theta_1), \ldots, \sigma_n(\theta_n))$, and the message profile sent by agents other than i in state $\theta = (\theta_{-i}, \theta_i)$ is denoted

$$\sigma_{-i}(\theta_{-i}) = (\sigma_1(\theta_1), \ldots, \sigma_{i-1}(\theta_{i-1}), \sigma_{i+1}(\theta_{i+1}), \ldots, \sigma_n(\theta_n)).$$

Let Σ denote the set of all strategy profiles. Strategy profile $\sigma \in \Sigma$ is a *Bayesian Nash Equilibrium* if and only if for all $i \in N$ and all $\theta_i \in \Theta_i$,

$$E\{u_i(h(\sigma(\theta_{-i}, \theta_i)), \theta_i) \mid \theta_i\} \geqslant E\{u_i(h(\sigma_{-i}(\theta_{-i}), m_i'), \theta_i) \mid \theta_i\},$$

for all $m_i' \in M_i$. All expectations are with respect to θ_{-i} conditional on θ_i. Let $\mathrm{BNE}^{\Gamma} \subseteq \Sigma$ denote the set of Bayesian Nash Equilibria for mechanism Γ.

A *social choice set* (SCS) is a collection $\widehat{F} = \{f_1, f_2, \ldots\}$ of social choice functions, i.e., a subset of A^{Θ}. We identify the SCF f: $\Theta \rightarrow A$ with the SCS $\widehat{F} = \{f\}$. Define the composition $h \circ \sigma$: $\Theta \rightarrow A$ by $(h \circ \sigma)(\theta) = h(\sigma(\theta))$. A mechanism $\Gamma = \langle M, h\rangle$ implements the SCS $\widehat{F}$ in Bayesian Nash equilibria if and only if (i) for all $f \in \widehat{F}$ there is $\sigma \in \mathrm{BNE}^{\Gamma}$ such that $h \circ \sigma = f$, and (ii) for all $\sigma \in \mathrm{BNE}^{\Gamma}$ there is $f \in \widehat{F}$ such that $h \circ \sigma = f$.

5.2. Closure

A set $\Theta' \subseteq \Theta$ is a *common knowledge event* if and only if $\theta' = (\theta_{-i}', \theta_i') \in \Theta'$ and $\theta = (\theta_{-i}, \theta_i) \notin \Theta'$ implies, for all $i \in N$, $p(\theta_{-i} \mid \theta_i') = 0$. If an agent is not sure about the true state, then in order to know what message to send he must predict what messages the other agents would send in all those states that he thinks are possible, which links a number of states together. However, two disjoint common knowledge events Θ_1 and Θ_2 are not linked in this way. For this reason, a necessary condition for Bayesian Nash implementation of an SCS $\widehat{F}$ is *closure* [Postlewaite and Schmeidler (1986), Palfrey and Srivastava (1989a), Jackson (1991)]: for any two common knowledge events Θ_1 and Θ_2 that partition Θ, and any pair $f_1, f_2 \in \widehat{F}$, we have $f \in \widehat{F}$ where f is defined by $f(\theta) = f_1(\theta)$ if $\theta \in \Theta_1$ and $f(\theta) = f_2(\theta)$ if $\theta \in \Theta_2$.

If every state is a common knowledge event, then we are in effect back to the case of complete information, and any SCS which satisfies closure is equivalent to an SCR. For an example of an SCS which does not satisfy closure, suppose $\Theta = \{\theta, \theta'\}$

where each state is a common knowledge event. The SCS is $\widehat{F} = \{f_1, f_2\}$, where $f_1(\theta) = f_2(\theta') = a$, $f_1(\theta') = f_2(\theta) = b$, and $a \neq b$. This SCS cannot be implemented. Indeed, to implement $\widehat{F}$ we would in effect need both a and b to be Nash equilibrium outcomes in both states, but then there would be no way to guarantee that the outcomes in the two states are different, as required by both f_1 and f_2. Notice that $\widehat{F}$ is not equivalent to the constant SCR F defined by $F(\theta) = F(\theta') = \{a, b\}$, since F does not incorporate the requirement that there be a different outcome in the two states.

5.3. *Incentive compatibility*

An SCF f is *incentive compatible* if and only if for all $i \in N$ and all $\theta_i, \theta_i' \in \Theta_i$,

$$E\{u_i(f(\theta_{-i}, \theta_i), \theta_i) \mid \theta_i\} \geqslant E\{u_i(f(\theta_{-i}, \theta_i'), \theta_i) \mid \theta_i\}.$$

An SCS $\widehat{F}$ is incentive compatible if and only if each $f \in \widehat{F}$ is incentive compatible[47].

Theorem 22: [Dasgupta, Hammond and Maskin (1979), Myerson (1979), Harris and Townsend (1981)]. *If the SCS $\widehat{F}$ is implementable in Bayesian Nash equilibria, then $\widehat{F}$ is incentive compatible.*

Proof: Suppose $\Gamma = \langle M, h\rangle$ implements $\widehat{F}$, but some $f \in \widehat{F}$ is not incentive compatible. Then there is $i \in N$ and $\theta_i, \theta_i' \in \Theta_i$ such that

$$E\{u_i(f(\theta), \theta_i) \mid \theta_i\} < E\{u_i(f(\theta_{-i}, \theta_i'), \theta_i) \mid \theta_i\}, \tag{8}$$

where $\theta = (\theta_{-i}, \theta_i)$. Let $\sigma \in \mathrm{BNE}^{\Gamma}$ be such that $h \circ \sigma = f$. If agent i's type θ_i uses the equilibrium strategy $\sigma_i(\theta_i)$, his expected payoff is

$$E\{u_i(h(\sigma(\theta)), \theta_i) \mid \theta_i\} = E\{u_i(f(\theta), \theta_i) \mid \theta_i\}. \tag{9}$$

If instead he were to send the message $m_i' = \sigma_i(\theta_i')$, he would get

$$E\{u_i(h(\sigma_{-i}(\theta_{-i}), \sigma_i(\theta_i'))) \mid \theta_i\} = E\{u_i(f(\theta_{-i}, \theta_i'), \theta_i) \mid \theta_i\}. \tag{10}$$

But Inequality (8) and Inequalities (9) and (10) contradict the definition of Bayesian Nash equilibrium. □

A mechanism Γ is a *revelation mechanism* if each agent's message is an announcement of his own type: $M_i = \Theta_i$ for all $i \in N$. Theorem 22 implies the *revelation principle*: if $\widehat{F}$ is implementable, then for each $f \in \widehat{F}$, truth telling is a Bayesian Nash equilibrium

[47] The terminology *Bayesian incentive compatibility* may be used to distinguished this condition from *dominant-strategy* incentive compatibility (strategy-proofness).

for the revelation mechanism $\langle M, h\rangle$ where $M_i = \Theta_i$ for each $i \in N$ and $h = f$. However, the revelation mechanism will in general have untruthful Bayesian Nash equilibria and will therefore not fully implement f [Postlewaite and Schmeidler (1986), Repullo (1986)].

5.4. Bayesian monotonicity

A *deception for agent i* is a function α_i: $\Theta_i \to \Theta_i$. A *deception* $\alpha = (\alpha_1, \ldots, \alpha_n)$ consists of a deception α_i for each agent i. Let $\alpha(\theta) \equiv (\alpha_1(\theta_1), \ldots, \alpha_n(\theta_n))$ and $\alpha_{-i}(\theta_{-i}) \equiv (\alpha_1(\theta_1), \ldots, \alpha_{i-1}(\theta_{i-1}), \alpha_{i+1}(\theta_{i+1}), \ldots, \alpha_n(\theta_n))$. The following definition is due to Jackson (1991), and is slightly stronger than the version given by Palfrey and Srivastava (1989a)[48].

Definition. *Bayesian monotonicity*: For all $f \in \widehat{F}$ and all deceptions α such that $f \circ \alpha \notin \widehat{F}$, there exist $i \in N$ and a function y: $\Theta_{-i} \to A$ such that

$$E\{u_i(f(\theta_{-i}, \theta_i), \theta_i) \mid \theta_i\} \geqslant E\{u_i(y(\theta_{-i}), \theta_i) \mid \theta_i\}, \tag{11}$$

for all $\theta_i \in \Theta_i$ and

$$E\{u_i(f(\alpha(\theta_{-i}, \theta_i')), \theta_i') \mid \theta_i'\} < E\{u_i(y(\alpha_{-i}(\theta_{-i})), \theta_i') \mid \theta_i'\}, \tag{12}$$

for some $\theta_i' \in \Theta_i$.

When agents have complete information, monotonicity guarantees that a mechanism can be built which has no undesirable Nash equilibria (compare the discussion in the first paragraph of Section 4.6). As the proof of Theorem 23 will make clear, Bayesian monotonicity plays exactly the same role in incomplete information environments. A related condition called *selective elimination* was introduced by Mookherjee and Reichelstein (1990).

Theorem 23: [Postlewaite and Schmeidler (1986), Palfrey and Srivastava (1989a), Jackson (1991)]. *If the SCS $\widehat{F}$ is implementable in Bayesian Nash equilibria, then $\widehat{F}$ is Bayesian monotonic.*

Proof: Suppose a mechanism $\Gamma = \langle M, h\rangle$ implements $\widehat{F}$ in Bayesian Nash equilibria. For each $f \in \widehat{F}$ there is $\sigma \in \mathrm{BNE}^{\Gamma}$ such that $h \circ \sigma = f$. Let α be a deception such that

[48] Palfrey and Srivastava (1989a) considered a different model of incomplete information. In their model, each agent observes an event (a set of states containing the true state). A set of events are *compatible* if they have non-empty intersection. Social choice functions only recommend outcomes for situations where the agents have observed compatible events. The social planner can respond to incompatible reports any way she wants, which (at least in economic environments) makes it easy to deter the agents from sending incompatible reports. Thus, Palfrey and Srivastava (1989a) found it sufficient to restrict their monotonicity condition to "compatible deceptions".

$f \circ \alpha \notin \widehat{F}$. Now, $\sigma \circ \alpha \in \Sigma$ is a strategy profile such that in state $\theta \in \Theta$ the agents behave as they would under σ if their types were $\alpha(\theta)$, i.e., they send message profile $(\sigma \circ \alpha)(\theta) = \sigma(\alpha(\theta))$. Since $h \circ (\sigma \circ \alpha) = f \circ \alpha \notin \widehat{F}$, implementation requires that $\sigma \circ \alpha \notin \mathrm{BNE}^{\Gamma}$. If $\sigma \circ \alpha$ is not a Bayesian Nash equilibrium, then some type $\theta_i' \in \Theta_i$ prefers to deviate to some message $m_i' \in M_i$. That is,

$$E\left\{u_i\left(h\left(\sigma\left(\alpha\left(\theta_{-i}, \theta_i'\right)\right)\right), \theta_i'\right) \mid \theta_i'\right\} < E\left\{u_i\left(h\left(\sigma_{-i}\left(\alpha_{-i}\left(\theta_{-i}\right)\right), m_i'\right), \theta_i'\right) \mid \theta_i'\right\}. \tag{13}$$

Let y: $\Theta_{-i} \to A$ be defined by $y(\theta_{-i}) = h(\sigma_{-i}(\theta_{-i}), m_i')$. Note that

$$y\left(\alpha_{-i}\left(\theta_{-i}\right)\right) = h\left(\sigma_{-i}\left(\alpha_{-i}\left(\theta_{-i}\right)\right), m_i'\right).$$

Now Inequality (12) follows from Inequality (13). Moreover, Inequality (11) must hold for each type $\theta_i \in \Theta_i$ by definition of Bayesian Nash equilibrium: when σ is played, each type $\theta_i \in \Theta_i$ prefers to send message $\sigma_i(\theta_i)$ rather than deviating to m_i'. □

Thus, the three conditions of closure, Bayesian monotonicity and incentive compatibility are necessary for Bayesian Nash implementation. Conversely, Jackson (1991) showed that in economic environments with $n \geqslant 3$, any SCS satisfying these three conditions can be Bayesian Nash-implemented. This improved on two earlier results for economic environments with $n \geqslant 3$: Postlewaite and Schmeidler (1986) proved the sufficiency of closure and Bayesian monotonicity when information is *non-exclusive*[49], and Palfrey and Srivastava (1989a) proved the sufficiency of closure together with their version of Bayesian monotonicity and a stronger incentive compatibility condition. For general environments with $n \geqslant 3$, Jackson (1991) shows that closure, incentive compatibility, and a condition called monotonicity-no-veto together are sufficient for Bayesian Nash implementation. The monotonicity-no-veto condition combines Bayesian monotonicity with no veto power. Dutta and Sen (1994) give an example of a Bayesian Nash-implementable SCF which violates monotonicity-no-veto. Even though there are only two alternatives and two possible types for each agent, any mechanism which implements their SCF must have an infinite number of messages for each agent.

Matsushima (1993) has shown that Bayesian monotonicity is a very weak condition if utility functions are quasi-linear and lotteries are available. In other environments, refinements can enlarge the set of implementable social choice functions.

Palfrey and Srivastava (1989b) showed that any incentive compatible SCF can be implemented in *undominated Bayesian Nash equilibria* if $n \geqslant 3$, value distinction and a full support assumption hold, and no agent is ever indifferent across all alternatives. For virtual Bayesian implementation see Abreu and Matsushima (1990), Duggan

[49] Information is non-exclusive if each agent's information can be inferred by pooling the other $n - 1$ agents' information. Palfrey and Srivastava (1987) discuss the implementability of well-known SCRs when information is non-exclusive.

(1997) and Serrano and Vohra (2001). For Bayesian implementation using sequential mechanisms see Baliga (1999), Bergin and Sen (1998) and Brusco (1995).

5.5. Non-parametric, robust and fault tolerant implementation

Most of the literature on Bayesian implementation assumes that the social planner who designs the mechanism knows the agents' common prior p. If she does not have this information, then the mechanism must be *non-parametric* in the sense that it cannot depend directly on p. However, the planner may be able to extract information about p by adding a stage where the agents report their beliefs. Choi and Kim (1999) construct such a mechanism for implementation in undominated Bayesian Nash equilibrium. They assume the agents' types are independently drawn from a distribution which is known to the agents but not to the social planner. In equilibrium, each agent truthfully reports his own beliefs as well as the beliefs of a "neighbor". Duggan and Roberts (1997) assume the social planner makes a prior point estimate of p, but implementation is required to be robust against small errors in this estimate.

A different kind of robustness was introduced by Corchón and Ortuño-Ortin (1995), who assumed agents are divided into local communities, each with at least three members. The social planner knows that information is complete within a community, but she does not necessarily know what agents in one community know about members of other communities. Implementation should be robust against different possible inter-community information structures. Yamato (1994) showed that an SCR is robustly implementable in this sense if and only if it is Nash-implementable.

Eliaz (2000) introduced *fault tolerant* implementation. The idea is that mechanisms ought not to break down if there are a few "faulty" agents who do not understand the rules of the game or make mistakes. Neither the social planner nor the (non-faulty) agents know which agent (if any) is faulty, but all other aspects of the true state are known to the (non-faulty) agents. A Nash equilibrium is *k-fault tolerant* if it is robust against deviations by at most k faulty players. When $n > 2(k + 1)$, any SCR that satisfies no veto power and a condition called k-monotonicity can be implemented in a fault tolerant way.

6. Concluding remarks

Many of the mechanisms exhibited in this survey are admittedly somewhat abstract and complicated. Indeed, implementation theory has sometimes been criticized for how different its mechanisms often seem from the simple allocation procedures – such as auctions – used in everyday life.

In our view, however, these criticisms are somewhat misplaced. The fundamental objective of this literature is to characterize which social choice rules are in principle implementable. In other words, the idea is to define the perimeter of the implementable set. Although considerations such as "simplicity" or "practicability" are undeniably

important, they will not even arise if the SCR in question is outside this set. Of course, once theoretical implementability has been established, the search for mechanisms with particular desirable properties can begin.

Relatedly, a major reason why many mechanisms in the implementation literature are so "complex" is that they are deliberately devised to work very generally. That is, they are constructed to implement a huge array of social choice rules in environments with little restriction. For example, the mechanism devised in the proof of Theorem 2 implements any monotonic SCR satisfying no-veto-power in a completely general social choice setting. Not surprisingly, one can ordinarily exploit the particular structure that derives from focusing on a particular SCR in a particular environment [a classic example is Schmeidler's (1980) simple implementation of the Walrasian rule in an economic environment].

In fact, we anticipate that, since so much has now been accomplished toward developing implementation theory at a general level, future efforts are likely to concentrate more on concrete applications of the theory, e.g., to contracts [see, for instance, Maskin and Tirole (1999)] or to externalities [see, for instance, Varian (1994)], where special structures will loom large.

Another direction in which we expect the literature to develop is that of bounded rationality. Most of implementation theory relies quite strongly on rationality: not only must agents be rational, but rationality must be common knowledge. It would be desirable to develop mechanisms that are more forgiving of at least limited departures from full-blown *homo game theoreticus*. The "fault tolerant" concept (see Section 5.5) developed by Eliaz (2000) is a step in that direction, but many other possible allowances for irrationalities could well be considered.

Finally, it would be worthwhile to allow for the possibility that agents have preferences over more than just the *outcomes* of a mechanism (i.e., that they care also about what transpires during the play of the mechanism). An interesting step along this line has been taken by Glazer and Rubinstein (1998).

References

Abreu, D., and H. Matsushima (1990), "Virtual implementation in iteratively undominated strategies: incomplete information", Mimeo (Princeton University).

Abreu, D., and H. Matsushima (1992a), "Virtual implementation in iteratively undominated strategies: complete information", Econometrica 60:993–1008.

Abreu, D., and H. Matsushima (1992b), "A response to Glazer and Rosenthal", Econometrica 60: 1439–1442.

Abreu, D., and H. Matsushima (1994), "Exact implementation", Journal of Economic Theory 64:1–19.

Abreu, D., and Arunava Sen (1990), "Subgame perfect implementation: a necessary and sufficient condition", Journal of Economic Theory 50:285–299.

Abreu, D., and Arunava Sen (1991), "Virtual implementation in Nash equilibria", Econometrica 59: 997–1022.

Arrow, K.J. (1963), Social Choice and Individual Values, 2nd Edition (Wiley, New York).

Baliga, S. (1999), "Implementation in incomplete information environments: the use of multi-stage games", Games and Economic Behavior 27:173–183.
Baliga, S., and S. Brusco (2000), "Collusion, renegotiation and implementation", Social Choice and Welfare 17:69–83.
Baliga, S., and T. Sjöström (1999), "Interactive implementation", Games and Economic Behavior 27: 38–63.
Baliga, S., L. Corchón and T. Sjöström (1997), "The theory of implementation when the planner is a player", Journal of Economic Theory 77:15–33.
Bergin, J., and Arunava Sen (1998), "Extensive form implementation in incomplete information environments", Journal of Economic Theory 80:222–256.
Black, D. (1958), The Theory of Committees and Elections (Cambridge University Press).
Bowen, H. (1943), "The interpretation of voting in the allocation of economic resources", Quarterly Journal of Economics 58:27–48.
Brusco, S. (1995), "Perfect Bayesian implementation", Economic Theory 5:429–444.
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.
Chakravorty, B. (1991), "Strategy space reduction for feasible implementation of Walrasian performance", Social Choice and Welfare 8:235–245.
Chakravorty, B., L. Corchón and S. Wilkie (1997), "Credible implementation", Games and Economic Behavior, forthcoming.
Choi, J., and T. Kim (1999), "A nonparametric, efficient public decision mechanism: undominated Bayesian Nash implementation", Games and Economic Behavior 27:64–85.
Clarke, E.H. (1971), "Multipart pricing of public goods," Public Choice 11:17–33.
Corchón, L. (1996), The Theory of Implementation of Socially Optimal Decisions in Economics (St. Martin's Press, New York).
Corchón, L., and I. Ortuño-Ortin (1995), "Robust implementation under alternative information structures", Economic Design 1:159–171.
Danilov, V. (1992), "Implementation via Nash equilibria", Econometrica 60:43–56.
Dasgupta, P., P.J. Hammond and E. Maskin (1979), "The implementation of social choice rules: some general results on incentive compatibility", Review of Economic Studies 46:185–216.
d'Aspremont, C., and L.A. Gérard-Varet (1979), "Incentives and incomplete information", Journal of Public Economics 11:25–45.
de Trenqualye, P. (1988), "Stability of the Groves and Ledyard mechanism", Journal of Economic Theory 46:164–171.
Deb, R. (1994), "Waiver, effectivity, and rights as game forms", Economica 61:167–178.
Deb, R., P.K. Pattanaik and L. Razzolini (1997), "Game forms, rights and the efficiency of social outcomes", Journal of Economic Theory 72:74–95.
Duggan, J. (1997), "Virtual Bayesian implementation", Econometrica 67:1175–1199.
Duggan, J., and J. Roberts (1997), "Robust implementation", Mimeo (University of Rochester, NY).
Dummett, M., and R. Farquharson (1961), "Stability in voting", Econometrica 29:33–43.
Dutta, B., and Arunava Sen (1991a), "Implementation under strong equilibria: a complete characterization", Journal of Mathematical Economics 20:49–67.
Dutta, B., and Arunava Sen (1991b), "A necessary and sufficient condition for two-person Nash implementation", Review of Economic Studies 58:121–128.
Dutta, B., and Arunava Sen (1994), "Bayesian implementation: the necessity of infinite mechanisms", Journal of Economic Theory 64:130–141.
Dutta, B., Arunava Sen and R. Vohra (1995), "Nash implementation through elementary mechanisms in economic environments", Economic Design 1:173–204.
Eliaz, K. (2000), "Fault tolerant implementation", Review of Economic Studies, forthcoming.

Farquharson, R. (1969), The Theory of Voting (Yale University Press).
Farrell, J. (1993), "Meaning and credibility in cheap-talk games", Games and Economic Behavior 5:514–531.
Fudenberg, D., and D. Levine (1998), The Theory of Learning in Games (MIT Press, Cambridge, MA).
Gaertner, W., P.K. Pattanaik and K. Suzumura (1992), "Individual rights revisited", Economica 59: 161–177.
Gärdenfors, P. (1981), "Rights, games and social choice", Nous 15:341–356.
Gaspart, F. (1996), "Fair implementation in the cooperative production problem: two properties of normal form mechanisms", Mimeo (FUND Namur, Belgium).
Gaspart, F. (1997), "A general concept of procedural fairness for one-stage implementation", Mimeo (FUND Namur, Belgium).
Gibbard, A.F. (1973), "Manipulation of voting schemes: a general result", Econometrica 41:587–601.
Glazer, J., and R. Rosenthal (1992), "A note on Abreu–Matsushima mechanisms", Econometrica 60: 1435–1438.
Glazer, J., and A. Rubinstein (1996), "An extensive game as a guide for solving a normal game", Journal of Economic Theory 70:32–42.
Glazer, J., and A. Rubinstein (1998), "Motives and implementation: on the design of mechanisms to elicit options", Journal of Economic Theory 79:157–173.
Green, J., and J.J. Laffont (1979), Incentives in Public Decision Making (North-Holland, Amsterdam).
Grossman, S., and M. Perry (1986), "Perfect sequential equilibrium", Journal of Economic Theory 39:97–119.
Groves, T. (1970), "The allocation of resources under uncertainty", Ph.D. dissertation (University of California, Berkeley).
Groves, T., and J. Ledyard (1977), "Optimal allocation of public goods: a solution to the 'free rider' dilemma", Econometrica 45:783–811.
Groves, T., and J. Ledyard (1987), "Incentive compatibility since 1972", in: T. Groves, R. Radner and S. Reiter, eds., Information, Incentives and Economic Mechanisms (University of Minnesota Press) pp. 48–111.
Groves, T., and M. Loeb (1975), "Incentives and Public Inputs", Journal of Public Economics 4:311–326.
Hammond, P.J. (1997), "Game forms versus social choice rules as models of rights", in: K.J. Arrow, A.K. Sen and K. Suzumura, eds., Social Choice Re-examined, Vol. 2 (Macmillan, London) pp. 82–95.
Harris, M., and R. Townsend (1981), "Resource allocation with asymmetric information", Econometrica 49:33–64.
Harsanyi, J.C., and R. Selten (1988), A General Theory of Equilibrium Selection in Games (MIT Press, Cambridge, MA).
Hayek, F. (1945), "The use of knowledge in society", American Economic Review 35:519–530.
Herrero, M., and S. Srivastava (1992), "Implementation via backward induction", Journal of Economic Theory 56:70–88.
Hong, L. (1995), "Nash implementation in production economies", Economic Theory 5:401–418.
Hurwicz, L. (1960), "Optimality and informational efficiency in resource allocation processes", in: K.J. Arrow, S. Karlin and P. Suppes, eds., Mathematical Methods in the Social Sciences (Stanford University Press) pp. 27–46.
Hurwicz, L. (1972), "On informationally decentralized systems", in: R. Radner and C.B. McGuire, eds., Decision and Organization (North-Holland, Amsterdam) pp. 297–336.
Hurwicz, L. (1977), "On the dimensional requirements of informationally decentralized processes", in: K.J. Arrow and L. Hurwicz, eds., Studies in Resource Allocation Processes (Cambridge University Press) pp. 413–424.
Hurwicz, L. (1979a), "Outcome functions yielding Walrasian and Lindahl allocations at Nash equilibrium points", Review of Economic Studies 46:217–225.
Hurwicz, L. (1979b), "On allocations attainable through Nash equilibria", Journal of Economic Theory 21:140–165.

Hurwicz, L., and D. Schmeidler (1978), "Construction of outcome functions guaranteeing existence and Pareto-optimality of Nash equilibria", Econometrica 46:1447–1474.

Hurwicz, L., and M. Walker (1990), "On the generic nonoptimality of dominant-strategy allocation mechanisms: a general theorem that includes pure exchange economies", Econometrica 58:683–704.

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: J. Ledyard, ed., The Economics of Informational Decentralization: Complexity, Efficiency and Stability (Kluwer Academic Publishers, Amsterdam) pp. 367–433.

Jackson, M. (1991), "Bayesian implementation", Econometrica 59:461–477.

Jackson, M. (1992), "Implementation in undominated strategies: a look at bounded mechanisms", Review of Economic Studies 59:757–775.

Jackson, M. (2001), "A crash course in implementation theory", Social Choice and Welfare 18:655–708.

Jackson, M., and T. Palfrey (2001), "Voluntary implementation", Journal of Economic Theory 98:1–25.

Jackson, M., T. 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.

Ledyard, J., and J. Roberts (1974), "On the incentive problem with public goods", Discussion Paper 116 (Center for Mathematical Studies in Economics and Management Science, Northwestern University).

Maskin, E. (1979a), "Incentive schemes immune to group manipulation", Mimeo (MIT, Cambridge, MA).

Maskin, E. (1979b), "Implementation and strong Nash equilibrium", in: J.J. Laffont, ed., Aggregation and Revelation of Preferences (North-Holland, Amsterdam) pp. 433–440.

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) pp. 173–204.

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 66:39–56.

Maskin, E., and J. Tirole (1999), "Unforeseen contingencies and incomplete contracts", Review of Economic Studies 66:83–114.

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 39:107–121.

McKelvey, R.D. (1989), "Game forms for Nash implementation of general social choice correspondences", Social Choice and Welfare 6:139–156.

Mookherjee, D., and S. Reichelstein (1990), "Implementation via augmented revelation mechanisms", Review of Economic Studies 57:453–476.

Moore, J. (1992), "Implementation, contracts and renegotiation in environments with complete information", in: J.J. Laffont, ed., Advances in Economic Theory, Vol. 1 (Cambridge University Press) pp. 182–282.

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), "Dominance solvable voting schemes", Econometrica 47:1337–1352.

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

Moulin, H., and B. Peleg (1982), "Cores of effectivity functions and implementation theory", Journal of Mathematical Economics 10:115–145.

Mount, K., and S. Reiter (1974), "The informational size of message spaces", Journal of Economic Theory 8:161–192.

Muench, T., and M. Walker (1984), "Are Groves–Ledyard equilibria attainable?", Review of Economic Studies 50:393–396.
Muller, E., and M.A. Satterthwaite (1977), "The equivalence of strong positive association and strategy proofness", Journal of Economic Theory 14:412–418.
Myerson, R. (1979), "Incentive compatibility and the bargaining problem", Econometrica 47:61–74.
Palfrey, T. (1992), "Implementation in Bayesian equilibrium: the multiple equilibrium problem in mechanism design", in: J.J. Laffont, ed., Advances in Economic Theory, Vol. 1 (Cambridge University Press) pp. 283–323.
Palfrey, T. (2001), "Implementation theory", in: R. Aumann and S. Hart, eds., Handbook of Game Theory, Vol. 3 (North-Holland, Amsterdam).
Palfrey, T., and S. Srivastava (1987), "On Bayesian implementable allocations", Review of Economic Studies 54:193–208.
Palfrey, T., and S. Srivastava (1989a), "Implementation with incomplete information in exchange economies", Econometrica 57:115–134.
Palfrey, T., and S. Srivastava (1989b), "Mechanism design with incomplete information: a solution to the implementation problem", Journal of Political Economy 97:668–691.
Palfrey, T., and S. Srivastava (1991), "Nash-implementation using undominated strategies", Econometrica 59:479–501.
Peleg, B. (1984), Game Theoretic Analysis of Voting in Committees (Cambridge University Press).
Peleg, B. (1998), "Effectivity functions, game forms, games and rights", Social Choice and Welfare 15:67–80.
Postlewaite, A., and D. Schmeidler (1986), "Implementation in differential information economies", Journal of Economic Theory 39:14–33.
Postlewaite, A., and D. Wettstein (1989), "Continuous and feasible implementation", Review of Economic Studies 56:603–611.
Reichelstein, S., and S. Reiter (1988), "Game forms with minimal message spaces", Econometrica 56:661–692.
Repullo, R. (1986), "The revelation principle under complete and incomplete information", in: K. Binmore and P. Dasgupta, eds., Economic Organization as Games (Basil Blackwell, Oxford) pp. 179–195.
Repullo, R. (1987), "A simple proof of Maskin's theorem on Nash implementation", Social Choice and Welfare 4:39–41.
Roberts, K.W.S. (1979), "The characterization of implementable choice rules", in: J.J. Laffont, ed., Aggregation and Revelation of Preferences (North-Holland, Amsterdam) pp. 321–348.
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.
Saijo, T., Y. Tatamitani and T. Yamato (1996), "Toward natural implementation", International Economic Review 37:949–980.
Samuelson, P.A. (1954), "The pure theory of public expenditure", Review of Economics and Statistics 36:387–389.
Sato, F. (1981), "On the informational size of message spaces for resource allocation processes in economies with public goods", Journal of Economic Theory 24:48–69.
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.
Satterthwaite, M.A., and H. Sonnenschein (1981), "Strategy-proof allocation mechanisms at differentiable points", Review of Economic Studies 48:587–597.
Schmeidler, D. (1980), "Walrasian analysis via strategic outcome functions", Econometrica 48:1585–1594.

Segal, I., and M. Whinston (2002), "The Mirrlees approach to mechanism design with renegotiation", Econometrica 70:1–47.
Sen, A.K. (1970), Collective Choice and Social Welfare (Holden-Day, San Francisco).
Sen, Arunava (1995), "The implementation of social choice functions via social choice correspondences: a general formulation and a limit result", Social Choice and Welfare 12:277–292.
Serrano, R., and R. Vohra (2001), "Some limitations of virtual Bayesian implementation", Econometrica 69:785–792.
Sjöström, T. (1991), "On the necessary and sufficient conditions for Nash implementation", Social Choice and Welfare 8:333–340.
Sjöström, T. (1993), "Implementation in perfect equilibria", Social Choice and Welfare 10:97–106.
Sjöström, T. (1994), "Implementation in undominated Nash equilibria without using integer games", Games and Economic Behavior 6:502–511.
Sjöström, T. (1996a), "Implementation by demand mechanisms", Economic Design 1:343–354.
Sjöström, T. (1996b), "Credibility and renegotiation of outcome functions in economics", Japanese Economic Review 47:157–169.
Sjöström, T. (1999), "Undominated Nash implementation with collusion and renegotiation", Games and Economic Behavior 26:337–352.
Smith, V. (1979), "Incentive compatible experimental processes for the provision of public goods", in: V.L. Smith, ed., Research in Experimental Economics, Vol. 1 (JAI Press, Greenwich, Conn.) pp. 59–168.
Suh, S.C. (1995), "An algorithm for checking strong Nash-implementability", Journal of Mathematical Economics 25:109–122.
Suh, S.C. (1996), "Implementation with coalition formation: a complete characterization", Journal of Mathematical Economics 26:409–428.
Suh, S.C. (1997), "Double implementation in Nash and strong Nash equilibria", Social Choice and Welfare 14:439–447.
Thomson, W. (1979), "Comment on Hurwicz: On allocations attainable through Nash equilibria", in: J.J. Laffont, ed., Aggregation and Revelation of Preferences (North-Holland, Amsterdam) pp. 420–431.
Thomson, W. (1996), "Concepts of implementation", Japanese Economic Review 47:133–143.
Thomson, W. (1999), "Monotonic extensions on economic domains", Review of Economic Design 4:13–33.
Tian, G. (1989), "Implementation of the Lindahl correspondence by a single-valued, feasible, and continuous mechanism", Review of Economic Studies 56:613–621.
Varian, H.R. (1994), "A solution to the problem of externalities when agents are well-informed", American Economic Review 84:1278–1293.
Vartiainen, H. (1999), "Subgame perfect implementation: a full characterization", Mimeo (University of Helsinki).
Vickrey, W.S. (1960), "Utility, strategy and social decision rules", Quarterly Journal of Economics 74:507–535.
Vickrey, W.S. (1961), "Counterspeculation, auctions, and competitive sealed tenders", Journal of Finance 16:8–37.
Walker, M. (1980), "On the nonexistence of a dominant strategy mechanism for making optimal public decisions", Econometrica 48:1521–1540.
Walker, M. (1981), "A simple incentive compatible mechanism for attaining Lindahl allocations", Econometrica 49:65–73.
Williams, S. (1986), "Realization and Nash implementation: two aspects of mechanism design", Econometrica 54:139–151.
Yamato, T. (1992), "On Nash implementation of social choice correspondences", Games and Economic Behavior 4:484–492.

Yamato, T. (1993), "Double implementation in Nash and undominated Nash equilibrium", Journal of Economic Theory 59:311–323.
Yamato, T. (1994), "Equivalence of Nash-implementability and robust implementability with incomplete information", Social Choice and Welfare 11:289–303.
Yamato, T. (1999), "Nash implementation and double implementation: equivalence theorems", Journal of Mathematical Economics 31:215–238.
Yoshihara, N. (2000), "A characterization of natural and double implementation in production economies", Social Choice and Welfare 17:571–599.
Zhou, L. (1991), "Inefficiency of strategy-proof allocation mechanisms in pure exchange economies", Social Choice and Welfare 8:247–254.