DIVISION OF THE HUMANITIES AND SOCIAL SCIENCES
CALIFORNIA INSTITUTE OF TECHNOLOGY
PASADENA, CALIFORNIA 91125

IMPLEMENTATION THEORY

Thomas R. Palfrey

SOCIAL SCIENCE WORKING PAPER 912

September 1995

Implementation Theory

Thomas R. Palfrey

Abstract

This surveys the branch of implementation theory initiated by Maskin (1977). Results for both complete and incomplete information environments are covered.

JEL classification numbers: 025, 026

Key words: Implementation Theory, Mechanism Design, Game Theory, Social Choice

Implementation Theory*

Thomas R. Palfrey

1 Introduction

Implementation theory is an area of research in economic theory that rigorously investigates the correspondence between normative goals and institutions designed to achieve (implement) those goals. More precisely, given a normative goal or welfare criterion for a particular class of allocation problems (or domain of environments) it formally characterizes organizational mechanisms that will guarantee outcomes consistent with that goal, assuming the outcomes of any such mechanism arise from some specification of equilibrium behavior. The approaches to this problem to date lie in the general domain of game theory because, as a matter of definition in the implementation theory literature, an institution is modelled as a mechanism, which is essentially a non-cooperative game. Moreover, the specific models of equilibrium behavior are usually borrowed from game theory. Consequently, many of the same issues that intrigue game theorists are the focus of attention in implementation theory: How do results change if informational assumptions change? How do results depend on the equilibrium concept governing stable behavior in the game? How do results depend on the distribution of preferences of the players or the number of players? Also, many of the same issues that arise in social choice theory and welfare economics are also at the heart of implementation theory: Are some first-best welfare criteria unachievable? What is the constrained second-best solution? What is the general correspondence between normative axioms on social choice functions and the possibility of strategic manipulation, and how does this correspondence depend on the domain of environments? What is the correspondence between social choice functions and voting rules?

In order to limit this chapter to manageable proportions, attention will be mainly focussed on the part of implementation theory using non-cooperative equilibrium concepts that follows the seminal (unpublished) paper by Maskin (1977). This body of research has its roots in the foundational work on decentralization and economic design of Hayek,

*This is prepared for *Handbook of Game Theory, Vol. 3* (R. Aumann and S. Hart, eds.). I thank John Duggan, Matthew Jackson, and Sanjay Srivastava for helpful comments and many enlightening discussions on the subject of implementation theory. Suggestions from Robert Aumann, Sergiu Hart and two anonymous readers are also gratefully acknowledged. I also wish to thank the National Science Foundation for its finanical support.

Koopmans, Hurwicz, Reiter, Marschak, Radner, Vickrey and others that dates back more than half a century.

The limitation to this somewhat restricted subset of the enormous literature of implementation theory and mechanism design excludes three basic categories of results. The first category is implementation via dominant strategy equilibrium, perhaps more familiarly known as the characterization of strategyproof mechanisms. This research, mainly following the seminar work of Gibbard (1973) and Satterthwaite (1975) has close connections with social choice theory, and for that reason has already been treated in some depth in Chapter 31 of this Handbook (Vol. 2). The second category excluded is implementation via solution concepts that allow for coalitional behavior, most notably, strong equilibrium (Dutta and Sen 1991c) and coalition-proof equilibrium (Bernheim, Peleg, and Whinston 1987). The third category involves practical issues in the design of mechanisms. The is a vast literature identifying specific classes of mechanisms such as divide-and-choose, sequential majority voting, auctions, and so forth, and studying/characterizing the range of social choice rules that are implementable using such mechanisms. Most of these topics are already covered in some detail in other chapters of this Handbook.

Later in this chapter we discuss practical aspects of implementation, as it relates to the specific mechanisms used in the constructive proofs of the main theorems. But a few introductory remarks about this are probably in order. In contrast to the literature devoted to studying classes of "natural" mechanisms, the most general characterizations of implementable social choice rules often resort to highly abstract mechanisms, with little or no concern for practical application. The reason for this is that the mechanisms constructed in these theorems are supposed to apply to arbitrary implementable social choice rules. A typical characterization result in implementation theory takes as given an equilibrium concept (say, subgame perfect Nash equilibrium) and tries to identify necessary and sufficient conditions for a social choice rule to be implementable, under minimal domain restrictions on the environment. The method of proof (at least for sufficiency) is to construct a mechanism that will work for *any* implementable social choice rule. That is, *a single game form* is identified which will implement all such rules, under the specified equilibrium concept. It should come as no surprise that this often involves the construction of highly abstract mechanisms.

A premise of the research covered in this chapter is that to create a general foundation for implementation theory it makes sense to begin by identifying general conditions on social choice rules (and domains and equilibrium concepts) under which there *exists* some implementing mechanism. In this sense it is appropriate to view the highly abstract mechanisms more as vehicles for proving existence theorems than as specific suggestions for the nitty gritty details of organizational design. This is not to say that they provide no insights into the principles of such design, but rather to say that for most practical applications one could (hopefully) strip away a lot of the complexity of the abstract mechanisms. Thus a natural next direction to pursue, particularly for those interested in specific applications, is the identification of practical restrictions on mechanisms and the characterization of social choice rules that can be implemented by mechanisms satisfying

such restrictions. Some research in implementation theory is beginning to move in this direction, and, by doing so, is beginning to bridge the gap between this literature and the aforementioned research which focuses on implementation using very special classes of procedures such as binary voting trees and bargaining rules.

1.1 The basic structure of the implementation problem

The simple structure of the general implementation problem provides a convenient way to organize the research that has been conducted to date in this area, and at the same time organize the possibilities for future research. Before, presenting this classification scheme, it is useful to present one of the most concise representations of the implementation problem, called a *Mount-Reiter diagram*, a version of which originally appeared in Mount and Reiter (1974).

[FIGURE 1 HERE]

This figure contains the basic elements of an implementation problem. The notation is as follows:

$\mathcal{E}$: The domain of environments. Each *environment*, $\in \mathcal{E}$, consists of a set of *feasible outcomes*, $A(e)$, a set of *individuals*, $I(e)$, and a *preference profile* $R(e)$, where $R_i(e)$ is a weak order on $A(e)$.[1]

X: The *outcome space.* Note: $A(e) \subseteq X$ for all $e \in \mathcal{E}$.

$F \subseteq \{f : \mathcal{E} \to X\}$: The *welfare criterion* (or *social choice set*) which specifies the set of acceptable mappings from environments to outcomes. An element, f, of F is called a *social choice function.*

$M = M_1 \times \ldots \times M_N$ The *message space.*

$g : M \to X$ The *outcome function.*

$\mu =< M, g >$ The *mechanism.*

$\Sigma =$ The *equilibrium concept* that maps each μ into $\Sigma_\mu \subseteq \{\sigma : \mathcal{E} \to M\}$

As an example to illustrate what these different abstract concepts might be, consider the domain of pure exchange economies. Each element of the domain would consist of a set of traders, each of whom has an initial endowment, and the set of feasible outcomes would just be the set of all reallocations of the initial endowment. Many implementation

[1]Except where noted, we will assume that the set of feasible outcomes is a constant A that is independent of e, and the planner knows A. Furthermore, we will typically take the set of individuals $I = \{1, 2, \ldots, N\}$ as fixed.

results rely on domain restrictions, which in this illustration would involve standard assumptions such as strictly increasing, convex preferences, and so forth. The welfare criterion, or social choice set might consist of all social choice functions satisfying a list of conditions such as individual rationality, Pareto optimality, interiority, envy-freeness, and so forth. One common social choice set is the set of all selections from the Walrasian equilibrium correspondence. For this case the message space might be either an agent's entire preference mapping or perhaps his demand correspondence, and the "planner" would take on the role of the auctioneer. An example of an outcome function would be the allocations implied by a deterministic pricing rule (such as some selection from the set of market clearing prices), given the reported demands. Common equilibrium concepts employed in these settings are Nash equilibrium or dominant strategy equilibrium.

The arrows of the Mount-Reiter diagram indicate that the diagram has the commutative property that, under the equilibrium concept Σ, the set of desirable social choice functions defined by F correspond exactly to the outcomes that arise under the mechanism μ. That is $F = g \circ \Sigma_\mu$. When this happens, we say that "μ implements F via Σ in $\mathcal{E}$." Whenever there exists some mechanism such that that statement is true, we say "F is *implementable* via Σ in $\mathcal{E}$." Implementation theory then looks at the relationship between domains, equilibrium concepts, welfare criteria, and implementing mechanisms, and the various questions that may arise about this relationship. The remainder of this chapter summarizes a small part of what is known about this. Because this has been the main focus of the literature, the discussion here will concentrate primarily on *the existence question*: Under what conditions on F, Σ, and $\mathcal{E}$ does there exist a mechanism μ such that μ implements F via Σ in $\mathcal{E}$?

1.2 Informational issues

To this point, nothing has been said about what information can be used in the construction of a mechanism nor about what information the individuals have about $\mathcal{E}$. A common interpretation given to the implementation problem is that there is a mythical agent, called "the planner," who has normative goals and very limited information about the environment (typically one assumes that the planner only knows $\mathcal{E}$ and X). The planner then must elicit enough information from the individuals to implement outcomes in a manner consistent with those normative goals. This requires creating incentives for such information to be voluntarily and accurately provided by the individuals. Nearly all of the literature in implementation theory assumes that the details of the environment are not directly verifiable by the planner, even *ex post*.[2] Thus, implementation theory characterizes the limits of a planner's power in a society with decentralized information.

The information that the individuals have about $\mathcal{E}$ is also an important consideration. This information is best thought of as part of the description of the domain. The main information distinction that is usually made in the literature is between *complete*

[2]There are some results on auditing and other *ex post* verification procedures. See, for example, Townsend (1979), Chander and Wilde (1992) and the references they cite.

information and *incomplete information.* Complete information models assume that e is common knowledge among the individuals (but, of course, unknown to the planner). Incomplete information assumes that individuals have some private information. This is usually modelled in the Harsanyi (1967–68) tradition, by defining an environment as a profile of types, one for each individual, where a type indexes an individual's information about other individuals' types. In this manner, an environment (preference profile, set of feasible allocations, etc.) is uniquely defined for each possible type profile.

One branch of implementation theory addresses a somewhat different informational issue. Given a social choice correspondence and a domain of environments, how much information about the environment is minimally needed to determine which outcome should be selected, and how close to this level of minimal information gathering (informationally efficient) do different "natural" mechanisms come? In particular, this question has been asked in the domain of neoclassical pure exchange environments, where the answer is that the Walrasian market mechanism is informationally efficient. (Hurwicz, 1977, Mount and Reiter (1974)). With few exceptions that branch of implementation theory does not directly address questions of incentive compatibility of mechanisms. This chapter will not cover the contributions in that area.

1.3 Equilibrium: Incentive Compatibility and Uniqueness

In principle, the equilibrium concept Σ could be almost anything. It simply defines a systematic rule for mapping environments into messages for arbitrary mechanisms. However, nearly all work in implementation theory and mechanism design restricts attention to equilibrium concepts borrowed from noncooperative game theory, all of which require the *rational response property* in one form or another. That is, each individual, given their information, preferences, and a set of assumptions about how other individuals are behaving, and given a set of rules $< M, g >$, adopts a rational response, where rationality is usually based on maximization of expected utility.[3]

The requirement of implementation can be broken down into two components. The first component is *incentive compatibility*:[4] This is most transparent for the special case of social choice functions (i.e., F is a singleton). If a mechanism $< M, g >$ implements a social choice function f, it must be the case that there is an equilibrium strategy profile $\sigma : \mathcal{E} \to M$ such that $g \circ \sigma = f$. The second component is *uniqueness*: If a mechanism $< M, g >$ implements a social choice function f, it must be the case, for all social choice functions $h \notin f$, that there is *not* an equilibrium strategy profile σ' such that $g \circ \sigma' = h$.

[3]There are some exceptions, notably the use of maximim strategies (Thomson 1979) and undominated strategies (Jackson 1992), which do not require players to adopt expected utility maximizing responses to the strategies of the other players.

[4]This is sometimes referred to as *Truthful Implementability* (Dasgupta, Hammond, and Maskin 1979) because, in the framework where individual preferences and information are represented as "types," if the mechanism is *direct* in the sense that each individual is required to report their type (i.e., $M = T$), then the truthful strategy $\sigma(t) = t$ is an equilibrium of the direct mechanism $\mu =< T, f >$.

For the more general case of the implementation of a social choice set, F, these two components extend in the natural way. If a mechanism $< M, g >$ implements a social choice set F, it must be the case that, for each $f \in F$ there is an equilibrium strategy profile σ such that $g \circ \sigma = f$. If a mechanism $< M, g >$ implements a social choice set F, it must be the case, for all social choice functions $h \neq F$, that there is *not* an equilibrium strategy profile σ such that $g \circ \sigma = h$.

1.4 The Organization of this Chapter

The remainder of this paper is divided into three sections. Section II presents characterizations of implementation under conditions of complete information for several different equilibrium concepts. In particular, the relatively comprehensive characterization for *Nash implementation* (i.e., implementation in complete information domains via Nash equilibrium) is set out in considerable detail. The partial characterizations for *refined Nash implementation* (subgame perfect equilibrium, undominated Nash equilibrium, dominance solvable equilibrium, etc.) are then discussed. This part also describes the problem of implementation by games of perfect information, and a few results in the area, particularly with regard to "voting trees," are briefly discussed. Finally results for *virtual implementation* (both in Nash equilibrium and using refinements) are described, where social choice functions are only approximated as the equilibrium outcomes of mechanisms.

Section III explains how the results for complete information are extended to incomplete information "Bayesian" domains, where environments are represented as collections of Harsanyi type-profiles, and players are assumed to have well-defined common knowledge priors about the distribution of types.

Section IV discusses some "difficult" problems in the area of implementation theory that have either been ignored or studied in only the simplest settings. This includes dynamic issues such as renegotiation of mechanisms and dynamic allocation problems, considerations of simplicity, robustness, and bounded rationality, and issues of incomplete control by the planner over the mechanism (side games played by the agents, or preplay communication).

2 Implementation under conditions of complete information

By complete information, we mean that individual preferences and feasible alternatives are common knowledge among all the individuals. This does *not* mean that the planner knows these preferences. The planner is assumed to know only the set of possible individ-

ual preference profiles and the set of feasible allocations.[5] For this reason, we simplify the notation considerably for this section of the chapter. First, we represent the domain by $\boldsymbol{R}$, the set of possible preference profiles, with typical element $R = (R_1, R_2, \ldots, R_N)$, and the set of feasible alternatives is A. A social choice set can, without loss of generality, be represented as a correspondence F mapping $\mathbf{R}$ into subsets of A. We denote the image of F at R by $F(R)$.

2.1 Nash Implementation

Consider a mechanism $\mu =< M, g >$ and a profile R. The pair (μ, R) defines a N-player noncooperative game.

Definition 1 *A message profile $m^* \in M$ is called a* Nash equilibrium of μ at R *if, for all $i \in I$, and for all $m_i \in M_i$*

$$g(m^*)R_i g(m_i, m^*_{-i}).$$

Therefore, the condition of Nash implementation is simply:

Definition 2 *A social choice correspondence F is* Nash implementable in $\mathbf{R}$ *if there exists a mechanism $\mu =< M, g >$ such that:*

1. *For every $R \in \boldsymbol{R}$ and for every $x \in F(\boldsymbol{R})$, there exists $m^* \in M$ such that m^* is a Nash equilibrium of μ at R and $g(m^*) = x$.*
2. *For every $R \in \boldsymbol{R}$ and for every $y \notin F(R)$, there does not exist $m^* \in M$ such that m^* is a Nash equilibrium of μ at R and $g(m^*) = y$.*

Alternatively, writing $\sigma^*(R)$ as the set of Nash equilibria of μ at R, we can state (a) and (b) as:

1. *$F(R) \subseteq g \circ \sigma^*(R)$ for all $R \in \boldsymbol{R}$*
2. *$g \circ \sigma^*(R) \subseteq F(R)$ for all $R \in$ boldmath R*

Condition (a) corresponds to what we have referred to as incentive compatibility and condition (b) is what we have referred to as uniqueness. We proceed from here by characterizing the implications of (a) and (b).

[5]Nearly always, the set of feasible allocations is taken as fixed in implementation theory. Notable exceptions to this are Hurwicz, Maskin, Postlewaite (1980), and Hong and Page (1994). In these papers, the mechanisms has individuals report endowments as well as preferences to the planner, but it is assumed that it is impossible for an individual to overstate his endowment (although understatements are possible). See Hong (1994) and Tian (1994) for extensions to Bayesian environments.

Incentive Compatibility

Suppose a social choice function $f : \boldsymbol{R} \to A$ is implementable via Nash equilibrium. Then there exists a mechanism μ that implements f via Σ in $\mathcal{E}$. What exactly does this mean? First, it means that there is a Nash equilibrium of μ that yields f as the equilibrium outcome function. With complete information this turns out to have very little bite. That is, examples of social choice functions that are not "incentive compatible" when the individuals have complete information are rather special. How special, you might ask. First, the examples must have only two individuals. This fact is quickly established below as Proposition 3.

Proposition 3 *In complete information domains with $N > 2$, every social choice function f is incentive compatible (I.e., there exists a mechanism such that (a) is satisfied.).*

Proof: Consider the following mechanism, which we call the *agreement mechanism.* Let $M_i = \boldsymbol{R}$ for all $i \in I$. That is, individuals report profiles.[6] Arbitrarily pick some $a_0 \in A$. Partition the message space into two parts, M_a (called the agreement region) and M_d (called the *disagreement region*).

$$M_a = \{m \in M | \exists j \in I, R \in \boldsymbol{R} \text{ such that } m_i = R \text{ for all } i \neq j\}$$

$$M_d = \{m \in M | m \notin M_a\}$$

In other words, the agreement region consists of all message profiles where either every individual reports the same preference profile, or all except one individual reports the same preference profile. The outcome function is then defined as follows.

$$\begin{aligned} g(m) = f(R) \quad & \text{if } m \in M_a \\ = a_o \quad & \text{if } m \in M_d \end{aligned}$$

It is easy to see that if the profile of preferences is R, then it is a Nash equilibrium for all individuals to report $m_i = R$, since unilateral deviations from unanimous agreement will not affect the outcome. Therefore, (a) is satisfied. ∎

Therefore, incentive compatibility[7] is not an issue when information is complete and $N > 2$. Of course the problem with this mechanism is that *any unanimous message*

[6] This mechanism could be simplified further by having each agent report a feasible outcome.

[7] The reader should not confuse this with a number of negative results on implementation of social choice functions via Nash equilibrium when mechanisms are restricted to being "direct" mechanisms ($M_i = R_j$). If individuals do not report profiles, but only report their own component of the profile (sometimes called *privacy preserving mechanisms*) then clearly incentive compatibility can be a problem. This kind of incentive compatibility is usually called strategyproofness, and is closely related to the problem of implementation under the much stronger equilibrium concept of dominant strategy equilibrium. See Dasgupta, Hammond, and Maskin (1979).

profile is a Nash equilibrium at any preference profile. Therefore, this mechanism does not satisfy the uniqueness requirement (b) of implementation. We return to this problem shortly, after addressing the question of incentive compatibility for the $N = 2$ case.

When $N = 2$ the outcome function of the mechanism used in Proposition 3 is not well defined, since a unilateral deviation from unanimous agreement is not well defined. If $m_1 = R$ and $m_2 = R'$, then it is unclear whether $g(m) = f(R)$ or $g(m) = f(R')$. There are some simple cases where incentive compatibility is assured when $N = 2$. First, if there exists a uniformly bad outcome, w, with the property that, for all a $\in A$, and for all $R \in$, a $R_i w, i = 1, 2$. In that case, the mechanism above can be modified so that M_a requires unanimous agreement, and $a_o = w$. Clearly any unanimous report of a profile is a Nash equilibrium regardless of the actual preferences of the individuals, so this modified mechanism satisfies (a) but fails to satisfy (b).

A considerably weaker assumption, called *nonempty lower intersection* is due to Dutta and Sen (1991b) and Moore and Repullo (1990). We state a slightly weaker version below, which is sufficient for the incentive compatibility requirement (a) when $N = 2$. They define a slightly stronger version that is needed to satisfy the uniqueness requirement (b).

Definition 4 *A social choice function f satisfies* Weak Nonempty Lower Intersection[8] *if, for all $R, R' \in \boldsymbol{R}$, such that $R \neq R'$, $\exists c \in A$ such that $f(R)R_1 c$ and $f(R')R'_2 c$.*

The definition of social choice correspondences is similar:

Definition 5 *A social choice correspondence F satisfies* Weak Nonempty Lower Intersection *if for all $R, R' \in \boldsymbol{R}$, such that $R \neq R'$, and for all $af(R)$ and $b \in f(R'), \exists c \in A$ such that $aR_1 c$ and $bR'_2 c$.*

To see that this is a sufficient condition for (a), consider implementing the social choice function f. From Definition 4, we can define a function $c(R, R')$ for $R \neq R'$ with the property that $f(R)R_1 c(R, R')$ and $f(R')R'_2 c(R, R')$. We can then modify the mechanism above by:

$$\begin{aligned} g(m) &= f(R) \quad \text{if } m_1 = m_2 = R \\ &= c(R, R') \quad \text{if } m_1 = R \text{and } m_2 = R' \end{aligned}$$

This mechanism is illustrated in Figure 2. It is easy to check that weak nonempty lower intersection guarantees $m = (R, R)$ is a Nash equilibrium when the actual profile is R.

FIGURE 2 HERE

[8]The stronger version, called Nonempty Lower Intersection, requires $f(R)P_1 c$ and $f(R')P'_2 c$.

There are two interesting special cases where Nonempty Lower Intersection holds. The first is when there exists a universally "bad" outcome (Moore and Repullo 1990) with the property that it is strictly less preferred than all outcomes in the range of the social choice rule, for all agents, at all profiles in the domain.[9] This is satisfied by any nonwasteful social choice rule in exchange economies with free disposal and strictly increasing preferences, since destruction of the endowment is a bad outcome. The second special case is any Pareto efficient and individually rational interior social choice correspondence in exchange economies (with or without free disposal) with strictly convex and strictly increasing preferences and fixed initial endowments (Dutta and Sen 1991b, Moore and Repullo 1990).

Uniqueness

Clearly, incentive compatibility places few restrictions on Nash implementable social choice functions (and correspondences) with complete information. The second requirement of uniqueness is more difficult, and the major breakthrough in characterizing this was the classic (unpublished) paper of Maskin (1977). In that paper, he introduces two conditions, which are jointly sufficient for Nash implementation when $N \geq 3$. These conditions are called *Monotonicity* and *No Veto Power (NVP).*

Definition 6 *A social choice correspondence F is* Monotonic *if, for all $R, R' \in R$*

$$(x \in F(R), x \notin F(R')) \Rightarrow \exists i \in I, a \in A \text{ such that } xR_iaP_i'x.$$

The agent i and the alternative a are called, respectively, the *test agent* and the *test alternative.* Stated in the contrapositive, this says simply that if x is a socially desired alternative at R, and x does not strictly fall in preference for anyone when the profile is changed to R', then x must be a socially desired alternative at R'. Thus monotonic social choice correspondences must satisfy a version of a nonnegative responsiveness criterion with respect to individual preferences. In fact, this is a remarkably strong requirement for a social choice correspondence. For example, it rules out nearly any scoring rule, such as the Borda count or Plurality voting. Several other examples of nonmonotonic social choice functions in applications to bilateral contracting are given in Moore and Repullo (1988). One very nice illustration of a nonmonotonic social choice correspondence is a variation on the "King Solomon's Dilemma" example of Glazer and Ma (1989) and Moore (1992). The problem is to allocate a baby to its true mother. There are two individuals in the game (Ms. α and Ms. β).

Example 7 Assume that there are four possible alternatives:

a = give the baby to Ms. α
b = give the baby to Ms. β
c = divide the baby into two equal halves and give each mother one half
d = execute both mothers and the child

[9] Notice that this is a joint restriction on the domain and the social choice function.

Also, assume the domain consists of only two possible preference profiles depending on whether α or β is the real mother, and we will call these profiles R and R' respectively. They are given below:

$$\begin{aligned} R^{\alpha} &= a \succ b \succ c \succ d \\ R'_{\alpha} &= a \succ c \succ b \succ d \\ R_{\beta} &= b \succ c \succ a \succ d \\ R'_{\beta} &= b \succ a \succ c \succ d \end{aligned}$$

The social choice function King Solomon wishes to implement is $f(R) = a$ and $f(R') = b$. This is not monotonic. Consider the change from R to R'. Alternative a does not fall in either player's preference order as a result of this change. However, $f(R') = b \neq a$, a contradiction of monotonicity. Notice however that this social choice function is incentive compatible since there is a universally bad outcome, d, which is ranked last by both players in both of their preference orders. $\triangle$

A second example, from a neoclassical 2-person pure exchange environment illustrates the geometry of monotonicity. Consider allocation x in figure 3.

[FIGURE 3 HERE]

Suppose $x \in f(R)$ where the indifference curves through x of the two individuals are labelled R_1 and R_2 respectively in that figure. Now consider some other profile R' where $R_2 = R'_2$, and R'_1 is such that the lower contour set of x for individual 1 has expanded. Monotonicity would require $x \in f(R')$. Put another way (formally stated in the definition) if f is monotonic and $x \notin f(R'')$ for some $R'' \neq R$, then one of the two individuals must have an indifference curve through x that either crosses the R-indifference curve through x or bounds a strictly smaller lower contour set. Figure 4 illustrates the (generic) case in which the R''-indifference curve of one of the individuals (individual 1, in the figure) crosses the R-indifference curve through x. Thus, in this example agent 1 is the test agent. One possible test alternative $a \in A$ (an alternative required in the definition of monotonicity which has the property that $xR_i aP''_i x$) is marked in that figure.

[FIGURE 4 HERE]

Maskin (1977) proved that monotonicity is a necessary condition for Nash implementation.

Theorem 8 *If F is Nash implementable then F is monotonic.*

Proof: Consider any mechanism μ that Nash implements F and consider some $x \in F(R)$ and some Nash equilibrium message, m^*, at profile R, such that $g(m^*) = x$. Define the "option set"[10] for i at m^* as

$$O_i(m^*; \mu) = \{a \in A | \exists m_i' \in M_i \text{ such that } g(m_i', m^*_{-i}) = a\}$$

That is, fixing the messages of the other players at m^*_{-i}, the range of possible outcomes that can result for some message by i under the mechanism μ is $O_i(ma^*; \mu)$. By the definition of Nash equilibrium, $m_i^* R_i$ a for all i and for all $a \in O_i(m^*; \mu)$. Now consider some new profile R' where $x \notin F(R')$. Since μ Nash implements F, it must be that m is not a Nash equilibrium at R'. Thus there exists some i and some alternative $a \in O_i(m^*; \mu)$ such that $aP_i'x$. Thus a is the test alternative and i is the test agent as required in Definition 6, with the property that $xR_iaP_i'x$. ∎

The second theorem in Maskin (1977), which was later given a complete proof by Williams (1984), Saijo (1988), McKelvey (1989), and Repullo (1987), provides a simple sufficient condition for Nash implementation for the case of three or more agents. This is a condition of near unanimity, called No Veto Power (NVP).

Definition 9 *A social choice correspondence F satisfies* No Veto Power (NVP) *if, for all $R \in \boldsymbol{R}$ and for all $x \in A$, and $i \in I$,*

$$[xR_jy \text{ for all } j \neq i, \text{ for all } y \in Y] \Rightarrow x \in F(R).$$

Theorem 10 *If $N \geq 3$ and F is Monotonic and satisfies NVP, then F is Nash implementable.*

Proof: (based on Repullo 1987) The proof is constructive, like the proof of Proposition 3. A very general mechanism is defined, and then the rest of the proof consists of demonstrating that, the mechanism implements any social choice function that satisfies the hypotheses of the theorem. This is usually how characterization theorems are proved in implementation theory. Consider the following generic mechanism, which we call the *agreement/integer mechanism*

$$M_i = \boldsymbol{R}A \times \{0, 1, 2, \ldots\}$$

That is, each individual reports a **profile**, an **allocation**, and an **integer**. The outcome function is similar to the agreement mechanism, except the disagreement region is a bit more complicated, and agreement must be with respect to an allocation and a profile.

[10]This is similar to the role of option sets in the strategyproofness literature.

$M_a = \{m \in M | \exists j, R \in \boldsymbol{R} a \in F(R)$ such that $m_i = (R, a, z_i)$ where $z_i = 0$ for each $i \neq j\}$

$M_d = \{m \in M | m \notin M_a\}$

The outcome function is defined as follows. The outcome function is constructed so that, if the message profile is in M_a, then the outcome is either a or the allocation announced by individual j, which we will denote a_j. If the outcome is in M_d, then the outcome is a_k where k is the individual announcing the highest integer (ties broken by a predetermined rule). This feature of the mechanism has become commonly known as *an integer game* (although in actuality, it is only a piece of the original game). Formally,

$$\begin{aligned} g(m) &= a \quad \text{if } m \in M_a \text{ and } a_j P_j a \\ &= a_j \quad \text{if } m \in M_a \text{ and } a R_j a_j \end{aligned}$$

$$g(m) = a_k \quad \text{if } m \in M_d \text{ and } k = \max\{i \in I | z_i \geq z_j \text{ for all } j \in I\}$$

Recall that we must show that $F(R) \subseteq NE_{|mu}(R)$ and $NE_\mu(R) \subseteq F(R)$ for all R, where $NE_\mu(R) = \{a \in A | \exists m \in M$ such that $a = g(m) R_i g(m_i', m_{-i})$ for all $i \in I, m_i' \in M_i\}$ is the set of Nash equilibrium outcomes to μ at R.

1. $F(R) \subseteq NE_\mu(R)$

   At any R, and for any $a \in F(R)$, there is a Nash equilibrium in which all individuals report $m_i = (R, a, 0)$. Such a message lies in M_a and any unilateral deviation also lies in M_a. The only unilateral deviation that could change the outcome in a deviation in which some player j reports an alternative a_j such that $a R_j a_j$. Therefore, a is a R_j-maximal element of $O_j(m; \mu)$ for all $j \in I$, so $m = (R, a, 0)$ is a Nash equilibrium.

2. $NE_\mu(R) \subseteq F(R)$

   This is the more delicate part of the proof, and is the part that exploits Monotonicity and NVP. (Notice that part (1) of the proof above exploited only the assumption that $N \geq 3$.) Suppose that $m \in NE_\mu(R)$ and $g(m) = a \notin F(R)$. First notice that it cannot be the case that all individuals are reporting $(R', a, 0)$ where $a \in F(R')$ for some $R' \in \boldsymbol{R}$. This would put the outcome in M_a and Monotonicity guarantees the existence of some $j \in I, b \in A$ such that a $R_i' b P_i a$, so that player is better off changing to a message $(\cdot, b, \cdot)$ which changes the outcome from a to b. Thus $m_i \neq m_j$ for some i, j. Whenever this is the case, the option set for at least $N - 1$ of the agents is the entire alternative space, A. Since $a \notin F(R)$ and F satisfies NVP, it must be that there is at least one of these $N - 1$ agents, k, and some element $c \in A$ such that $c P_k a$. Since the option set for k is the entire alternative space, A, individual k is better off changing his message to $(\cdot, c, z_k) \neq m_j$ where $z_k > z_j$, $j \neq k$, which will change the outcome to from a to c. This contradicts the hypothesis that m is a Nash equilibrium.

∎

Since these two results, improvements have been developed to make the characterization of Nash implementation complete and/or to reduce the size of the message space of the general implementing mechanism. These improvements are in Moore and Repullo (1990), Dutta and Sen (1991b), Danilov (1992),[11] Saijo (1988), Sjöström (1991b) and McKelvey (1989) and the references they cite.

The last part of this section on Nash implementation is devoted to a simple application to pure exchange economies. It turns out the Walrasian correspondence satisfies both Monotonicity and NVP under some mild domain restrictions. First notice that in private good economies with strictly increasing preferences and three or more agents, NVP is satisfied vacuously. Next suppose that indifference curves are strictly quasi-concave and twice continuously differentiable, endowments for all individuals are strictly positive in every good, and indifference curves never touch the axis. It is well known that these conditions are sufficient to guarantee the existence of a Walrasian equilibrium and to further guarantee that all Walrasian equilibrium allocations in these environments are interior points, with every individual consuming a positive amount of every good in every competitive equilibrium. Finally, assume that "the planner" knows everyone's endowment.[12]

Since the planner knows the endowments, a different mechanism can be constructed for each endowment profile. Thus, to check for monotonicity it suffices to show that the Walrasian correspondence, with endowments fixed and only preferences changing, is monotonic. If a is a Walrasian equilibrium allocation at R and not a Walrasian equilibrium allocation at R', then there exists some individual for whom the supporting price line for the equilibrium at R is not tangent to the R_i' indifference curve through a. But this is just the same as the illustration in Figure 4, and we have labelled allocation b as a "test allocation" as required by the monotonicity definition. The key is that for a to be a Walrasian equilibrium allocation at R and not a Walrasian equilibrium allocation at R' implies that the indifference curves through x at R and R' cross at x.

As mentioned briefly above, there are many environments and "nice" (from a normative standpoint) allocation rules that violate Monotonicity, and in the $N = 2$ case ("bilateral contracting" environments) NVP is simply too strong a condition to impose on a social choice function. There are two possible responses to this problem. One possibility, and the main direction implementation theory has pursued, is that Nash

[11] Of these, Danilov (1992) establishes a particularly elegant necessary and sufficient condition (with three or more players), which is a generalization of the notion of monotonicity, called *essential monotonicity*. However, these results are limited somewhat by this assumption of universal domain. Nash implementable social choice correspondences need not satisfy essential monotonicity under domain restrictions.

[12] This assumption can be relaxed. See Hurwicz, Maskin, and Postlewaite (1980). The Walrasian correspondence can also be modified somewhat to the "constrained Walrasian correspondence" which constrains individual demands in a particular way. This modified competitive equilibrium can be shown to be implementable in more general economic domains in which Walrasian equilibrium allocations are not guaranteed to be locally unique and interior. See the survey by Postlewaite (1985), or Hurwicz (1986).

equilibrium places insufficient restrictions on the behavior of individuals.[13] This leads to consideration of implementation using refinements of Nash equilibrium, or *refined Nash implementation.* A second possibility is that implementability places very strong restrictions on what kinds of social choice functions a planner can hope to enforce in a decentralized way. If not all social choice functions can be implemented, then we need to ask "how close" can we get to implementing a desired social choice function? This has led to the work in *virtual implementation.* These two directions are discussed next.

2.2 Refined Nash implementation

More social choice correspondences can be implemented using refinements of Nash equilibrium. The reason for this is straightforward, and is easiest to grasp in the case of $N \geq 3$. In that case, the incentive compatibility problem does not arise (Proposition 3), so the only issue is (ii) uniqueness. Thus the problem with Nash implementation is that Nash equilibrium is too permissive an equilibrium concept. A nonmonotonic social choice function fails to be implementable simply because there are too many Nash equilibria. It is impossible to have $f(R)$ a Nash equilibrium outcome at R and at the same time avoid having $a \neq f(R)$ also be a Nash equilibrium outcome at R. But of course this is exactly the kind of problem that refinements of Nash equilibrium can be used for. The trick in implementation theory with refinements is to exploit the refinement by constructing a mechanism so that precisely the "bad" equilibria (the equilibria whose outcomes lie outside of F) are refined away, while the other equilibria survive the refinement.[14]

2.2.1 Subgame perfect implementation

The first systematic approach to extending the Maskin characterization beyond Nash equilibrium in complete information environments was to look at implementation via subgame perfect Nash equilibrium (Moore and Repullo (1988) and Abreu and Sen (1990)). They find that more social choice functions can be implemented via subgame perfect Nash equilibrium than via Nash equilibrium. The idea is that sequential rationality can be exploited to eliminate certain bad equilibria. The following simple example in the "voting/social choice" tradition illustrates the point.

[13] One might argue to the contrary that in other ways Nash equilibrium places too strong a restriction on individual behavior. Both directions are undoubltedly true. Experimental evidence has shown that both of these are defensible. On the one hand, some refinements of Nash equilibrium have received experimental support indicating that additional restrictions beyond mutual best response have predictive value (Banks, Camerer, and Porter (1994)). On the other hand, many experiments indicate that players are at best imperfectly rational, and even violate simple basic axioms such as transitivity and dominance. Thus, from a practical standpoint, it is very important to explore the implementation question under assumptions that other than the simple mutual best response criterion of Nash equilibrium.

[14] Earlier work by Farquharson (1957/69), Moulin (1979), Crawford (1979) and Demange (1984) in specific applications of multistage games to voting theory, bargaining theory, and exchange economies foreshadows the more abstract formulation in the relatively more recent work in implementation theory with refinements.

Example 11 There are three players on a committee who are to decide between three alternatives, $A = \{a, b, c\}$. There are two profiles in the domain, denoted R and R'. Individuals 1 and 2 have the same preferences in both profiles. Only player 3 has a different preference order under R than under R'. These are listed below:

$$R_1 = R'_1 = a \succ b \succ c$$
$$R_2 = R'_2 = b \succ c \succ a$$
$$R_3 = c \succ a \succ b$$
$$R'_3 = a \succ c \succ b$$

The following social choice function is Pareto optimal and satisfies the Condorcet criterion that an alternative should be selected if it preferred by a majority to any other alternative:

$$f(R) = b$$
$$f(R') = a$$

This social choice function violates monotonicity since b does not go down in player 3's rankings moving from profile R to R' (and no one else's preferences change). Therefore it is not Nash implementable. However, the following trivial mechanism (extensive form game form) implements it in subgame perfect Nash equilibrium:

Stage 1: Player 1 either chooses alternative b, or passes. The game ends if b is chosen. The game proceeds to stage 2 if player 1 passes.

Stage 2: Player 3 chooses between a and c. The game ends at this point.

The voting tree is illustrated in Figure 5.

[FIGURE 5 HERE]

To see that this game implements f, work back from the final stage. In stage 2, player 3 would choose c in profile R and a in profile R'. Therefore, player 1's best response is to choose b in profile R and to pass in profile R'. Notice that there is another Nash equilibrium under profile R', where player 2 adopts the strategy of choosing c if player 1 passes, and thus player 1 chooses b in stage 1. But of course this is not sequentially rational and is therefore ruled out by subgame perfection. $\triangle$

Abreu and Sen (1990) provide a nearly complete characterization of social choice correspondences that are implementable via subgame perfect Nash equilibrium, by giving a general necessary condition, which is also sufficient if $N \geq 3$ for social choice functions satisfying NVP. This condition is strictly weaker than Monotonicity, in the following way. Recall that monotonicity requires, for any R, R' and $a \in A$, with $a = f(R)$, $a \neq f(R')$, the

existence of a test agent, i and a test allocation b such that $aR_ibR_i'a$. That there is some player and some allocation that produces a preference switch with $f(R)$ when moving from R to R'. The weakening of this resulting from the sequential rationality refinement is that the preference switch does not have to involve $f(R)$ directly. Any preference switch between two alternatives, say b and c will do, as long as these alternative can be indirectly linked to $f(R)$ in a particular fashion. We formally state this necessary condition and call it *indirect monotonicity*,[15] to contrast it with the direct linkage to $f(R)$ of the test alternative in the original definition of monotonicity.

Definition 12 *A social choice correspondence F satisfies* indirect monotonicity *if there $\exists B \subseteq A$ such that $F(R) \subseteq B$ for all $R \in \mathcal{R}$, and if for all R, R' and $a \in A$, with $a \in F(R)$, $a \notin F(R'), \exists L < \infty$, and $\exists$ a sequence of agents $\{j_0, \ldots, j_L\}$ and $\exists$ sequences of alternatives $\{a_0, \ldots, a_{L+1}\}$, $\{b_0, \ldots, b_L\}$ belonging to B such that:*

$$(i) \quad a_k R_{jk} a_{k+1} \qquad k = 0, 1, \ldots, L$$

$$(ii) \quad a_{L+1} P') j L a_L$$

$$(iii) \quad b_k P'_{jk} a_k \qquad k = 0, 1, \ldots, L$$

$$(iv) \quad (a_{L+1} R' j_{Lb} \forall b \in B) \Rightarrow (L = 0 \text{ or } j_{L-1} = j_L)$$

The key parts the definition of indirect monotonicity are (i) and (ii). A less restrictive version of indirect monotonicity consisting of only parts (i) and (ii) was used first by Moore and Repullo (1988) as a weaker necessary condition for implementation in subgame perfect equilibrium with multistage games.

The main two general theorems about implementation via subgame perfect implementation are the following. The proofs (Abreu and Sen, 1990) are long and tedious and are omitted, although an outline of the proof for the sufficiency result is given. Similar results, but slightly less general, can be found in Moore and Repullo (1988).

Theorem 13 *(necessity) If a social choice correspondence F is implementable via subgame perfect Nash equilibrium, then F satisfies indirect monotonicity.*

Theorem 14 *(sufficiency) If $N \geq 3$ and F satisfies NVP and indirect monotonicity, then F is implementable via subgame perfect Nash equilibrium.*

Proof: Since F satisfies indirect monotonicity, there exists the required set B and for any (R, R', a) such that $a \in F(R)$ and $a \notin F(R')$ there exists an integer L and the required sequences $\{j_k(R, R', a)\})k = 0, 1, \ldots, L$ and $\{a_k(R, R', a)\}_{k=0,1,\ldots,L+1}$ that satisfy (i)–(iv) of Definition 6. In the first stage of the mechanism, all agents announce a triple of the form (m_{i1}, m_{i2}, m_{i3}) where $m_{i1} \in \boldsymbol{R}, m_{i2} \in A$, and $m_{i3} \in \{0, 1, \ldots\}$. The first stage of

[15] Abreu and Sen (1990) call it *Condition* α.

the game then conforms fairly closely to the agreement/integer mechanism, with a minor exception. If there is too much disagreement (there exist three or more agents whose reports are different) the outcome is determined by m_{i2} of the agent who announced the largest integer. If there is unanimous agreement in the first two components of the message, so all agent send some (R, a, z_i) and $a \in F(R)$, then the game is over and the outcome is a. The same is true if there is only one disagreeing report in the first two components, unless the dissenting report is sent by $i_0(R, m_{i01}, a)$, in which case the first of a sequence of at most L "binary" agreement/integer games is triggered in which either some agent gets to choose his most preferred element of B or the next in the sequence of binary agreement/integer games is triggered. If the game ever gets to the $(L+1)^{st}$ stage, then the outcome is a_{L+1} and the game ends.

The rest of the proof follows the usual order. First one shows that for all $R \in \mathcal{R}$ and for all $a \in F(R)$ there is a subgame perfect Nash equilibrium at R with a as the equilibrium outcome. Second one shows that for all $R \in \mathcal{R}$ and for all $a \notin F(R)$ there is no subgame perfect Nash equilibrium at R with a as the equilibrium outcome. ∎

In spite of its formidable complexity, some progress has been made tracing out the implications of indirect monotonicity for two well-known classes of implementation problems: exchange economies and voting. Moore and Repullo (1988) show that any selection from the Walrasian equilibrium correspondence satisfies indirect monotonicity, in spite of being nonmonotonic. There are also some results for the $N = 2$ case that can be found in Moore (1992) and Moore and Repullo (1988) which rely on sidepayments of a divisible private good. The case of voting-based social choice rules contrasts sharply with this. Abreu and Sen (1990), Palfrey and Srivastava (1991a) and Sen (1987) show that many voting rules fail to satisfy indirect monotonicity, as do most runoff procedures and "scoring rules" (such as the famous Borda rule). However, a class of voting-based social choice correspondences, including the Copeland rule, is implementable via subgame perfect Nash equilibrium (Sen 1987). Some related findings are in Moulin (1979), Dutta and Sen (1993), and the references they cite.

There are a number of applications that exploit the combined power of sidepayments and sequential mechanisms. See Glazer and Ma (1989), Varian (1993), and Jackson and Moulin (1990). Moore (1992) also gives some additional examples.

2.2.2 Implementation by backward induction and voting trees

In general, it is not possible to implement a social choice function via subgame perfect Nash equilibrium without resorting to games of imperfect information. At some point, it is necessary to have a stage with simultaneous moves. Others have investigated the implementation question when mechanisms are restricted to be games of perfect information. In that case, the refinement implied by solving the game in its last stage and working back to earlier moves, generates similar behavior as subgame perfect equilibrium.[16] Ex-

[16] In fact, it is exactly the same if players are assumed to have strict preferences. Much of the work in this area has evolved as a branch of social choice theory, where it is common to work with environments

ample 11, above illustrates how it is possible for nonmonotonic social choice functions to be implemented via backward induction. The work of Glazer and Ma (1989) illustrates how economic contracting and allocation problems similar in structure to example 7 (King Solomon's Dilemma) can be solved with backward induction implementation if sidepayments are possible. Crawford's work (1977, 1979) on bargaining mechanisms[17] proves in fairly general bargaining setting that games of perfect information can be used to implement nonmonotonic social choice functions that are fair. The general problem of implementation by backward induction has been studied by Hererro and Srivastava (1992) and Trick and Srivastava (1994). The characterizations, unfortunately, are quite cumbersome to deal with, and the necessary conditions for implementation via backward induction are virtually impossible to check in most settings. But some useful results have been found for certain domains.

Closely related to the problem of implementation by backward induction is implementation by voting trees, using the solution concept of *sophisticated voting* as developed by Farquharson (1957/69). Sophisticated voting works in the following way. First, a binary voting tree is defined, which consists of an initial pair of alternatives, which the individuals vote between. Depending on which outcome wins a majority of votes,[18] the process either ends or moves on to another, predetermined pair and another vote is taken. Usually one of the alternatives in this new vote is the winner of the previous vote, but this is not a requirement of voting trees. The tree is finite, so at some point the process ends regardless of which alternative wins. Sophisticated voting means that one starts at the end of the voting tree, and, for each "final" vote, determines who will win if everyone votes sincerely at that last stage. Then one moves back one step to examine every penultimate vote, and voters vote taking account of how the final votes will be determined. Thus, as in subgame perfect Nash equilibrium, voters have perfect foresight about the outcomes of future votes, and vote accordingly.

The problem of implementation by voting trees was first studied in depth by Moulin (1979), using the concept of dominance solvability, which reduces to sophisticated voting (McKelvey and Niemi 1978) in binary voting trees. There are two distinct types of sequential voting procedures that have been investigated in detail. The first type consists of *binary amendment procedures*. In a binary amendment procedure, all the alternatives (assumed to be finite) are place in a fixed order, say, $(a_1, a_2, \ldots, a_{|A|})$. At stage 1, the first two alternatives are voted between. Then the winner goes against the next alternative in the list, and so forth. A major question in social choice theory, and for that matter, in implementation theory, is to characterize the set of social choice functions that are implementable by binary amendment procedures via sophisticated voting. This work is closely related to work by Miller (1977), Banks (1985), and others, which explores general properties of the majority rule dominance relation, and following in the footsteps of Condorcet, looks at the implementability of social choice correspondences that satisfy

where A is finite and preferences are linear orders on A (i.e., strict.)

[17]This includes the divide-and-choose method and generalizations of it.

[18]It is common to assume an odd number of voters for obvious reasons. Extensions to permit even numbers of voters are usually possible, but the occurrence of ties clutters up the analysis.

certain normative properties. Several results appear in Moulin (1986), who identifies an "adjacency condition" that is necessary for implementation via binary voting trees. For more details, the reader is referred to the chapter of Social Choice Theory by Moulin (1993) in Volume 2 of this Handbook.

More recent results on implementation via binary voting trees are found in Dutta and Sen (1993). First, they show that implementability by sophisticated voting in binary voting trees implies implementability in backward induction using games of perfect information. The also show that several well-known selections from the top-cycle set[19] are implementable, but that certain selections that have appealing normative properties are not implementable.

2.2.3 Normal form refinements

There are some other refinements of Nash equilibrium that have been investigated for mechanisms in the Normal form. These fall into two categories. The first category relies on dominance (either strict or weak) to eliminate outcomes that are unwanted Nash equilibria. This was first explored in Palfrey and Srivastava (1991a) where implementation via *undominated Nash equilibrium* is characterized. Subsequent work that explores this and other variations of dominance-based implementation in the normal form includes Jackson (1992), Jackson, Palfrey, and Srivastava (1994), Sjöstrom (1991a), Tatamitami (1991) and Yamato (1993). Using a somewhat different approach, Abreu and Matsushima (1990) obtain results for implementation in iteratively weakly undominated strategies, if randomized mechanisms can be used and small fines can be imposed out of equilibrium. The work by Abreu and Matsushima (1992a, 1992b), Glazer and Rosenthal (1992), and Duggan (1993) extends this line of exploiting strategic dominance relations to refine equilibria by looking at iterated elimination of strictly dominated strategies and also investigating the use of these dominance arguments to design mechanisms that *virtually* implement (see Section 2.3 below) social choice functions.

The second category of refinements looks at implementation via trembling hand perfect Nash equilibrium. The main contribution here is the work of Sjöstrom (1993).

The central finding of the work in implementation theory using normal form refinements is that essentially anything can be implemented. In particular, it is the case that dominance-based refinements are more powerful than refinements based on sequential rationality, at least in the context of implementation theory. A simple result is in Palfrey and Srivastava (1991a), for the case of undominated Nash equilibrium.

Definition 15 *Consider a mechanism* $\mu =< M, g >$. *A message profile* $m^* \in M$ *is called an* undominated Nash equilibrium of μ at R *if, for all* $i \in I$, *for all* $m_i \in M_i$, $g(m^*)R_i g(m_i, m^*_{-i})$ *and there does not exist* $i \in I$ *and* $m_i \in M$ *such that*

[19]The top cycle set at R is the minimal subset, TC, of A, with the property that for all a, b such that $a \in TC$ and $b \notin TC$, a majority strictly perfer a to b. This set has a very prominent role in the theory of voting and committees.

$g(m_i, m_{-i}) R_i g(m_i^*, m_{-i})$ *for all* $m_{-i} \in M_{-i}$ *and*

$g(m_i, m_{-i}) P_i g(m_i^*, m_{-i})$ *for some* $m_{-i} \in M_{-i}$.

In other words, m^* *is an undominated Nash equilibrium at* R *if it is a Nash equilibrium and, for all* i, m_i^* *is not weakly dominated.*

Theorem 16 *Suppose* R *contains no profile where some agent is indifferent between all elements of* A. *If* $N \geq 3$ *and* F *satisfies NVP, then* F *is implementable in undominated Nash equilibrium.*

Proof: The proof of this theorem is quite involved. It uses a variation on the agreement/integer game, but the general construction of the mechanism uses an unusual technique, called tailchasing. Consider the standard agreement/integer game, μ, used in the proof of Theorem 10. If m^* is a Nash equilibrium of μ at R, but $g(m^*) \notin f(R)$, then one can make m^* dominated at R by amending the game in a simple way. Take some player i and two alternatives x, y such that xP_iy. Add a message m_i' for a player i and a message for each of the other players $j \neq i, m_j'$, such that

$$\begin{aligned} g(m_i^*, m_{-i}) &= g(m_i', m_{-i}) \text{ for all } m_{-i} \neq m'_{-i} \\ g(m_i^*, m'_{-i}) &= y \\ g(m_i', m'_{-i}) &= x \end{aligned}$$

Now strategy m^* is dominated at R. Of course, this is not the end of the story, since it is now possible that $(m'i, m^*_{-i})$ is a new undominated Nash equilibrium which still produces the undesired outcome $a \notin f(R)$. To avoid this, we add can another message for i, m_i'' and another message for the other players $j \neq i, m_j''$ and do the same thing again. If we repeat this an infinite number of times, we have created an infinite sequence of strategies for i, each one of which is dominated by the next one in the sequence. The complication in the proof is to show that in the process of doing this, we have not disturbed the "good" undominated Nash equilibria at R and have not inadvertently added some new undominated Nash equilibria. ∎

This kind of construction is illustrated in the following example.

Example 17 (from Palfrey and Srivastava 1991a, p. 488–89)

$$A = \{a, b, c, d\} \quad N = 2, \quad \boldsymbol{R} = \{R, R'\}$$

$$R_1 = R_2 \qquad \underline{R}'_1 = R'_2$$

a	a
b	b
cd	c
	d

$$F(R) = \{a, b\} \quad F(R') = \{a\}$$

It is easy to show that there is no implementation with a finite mechanism, and any implementation must involve an infinite chain of dominated strategies for one player in profile R'. One such mechanism is:

Player 2

		M_2^1	M_2^2	M_2^3	M_2^4	$\cdots$
	m_1^1	a	c	c	c	$\cdots\cdots$
Player	m_1^2	c	b	d	d	$\cdots\cdot$
1	m_1^3	c	b	c	d	$\cdots$
	m_1^4	c	b	c	c	$\cdots$
	$\cdot$	$\cdot$	$\cdot$	$\cdot$	$\cdot$	
	$\cdot$	$\cdot$	$\cdot$	$\cdot$	$\cdot$	
	$\cdot$	$\cdot$	$\cdot$	$\cdot$	$\cdot$	

Few would argue that mechanisms of this sort solve the implementation problem in a satisfactory manner.[20] This concern motivated the work of Jackson (1992) who raises the issue of *bounded* mechanisms.[21] △

[20] In fact, few would argue that *any* of the mechanisms used in the most general sufficiency theorems are particularly appealing.

[21] Boundedness is not the first property of mechanisms that has been investigated. Hurwicz (1960) suggests a number of criteria for judging the adequacy of a mechanism. Saijo (1988), McKelvey (1989), Dutta, Sen, and Vohra (1994), Reichelstein and Reiter (1988) and others have argued that message spaces should be as small as possible and have given results about how small the message spaces of implementing mechanisms can be. Abreu and Sen (1990) argue that mechanisms should have a best response property relative to the domain for which they are designed. Reichelstein (1984), Postlewaite and Wettstein (1989), and Wettstein (1992) analyze continuity of outcome functions as a property of implementing mechanisms.

Definition 18 *A mechanism is* bounded relative to $\boldsymbol{R}$ *if, for all* $R \in \boldsymbol{R}$, *and* $m_i \in M_i$, *if* m_i *is weakly dominated at* R, *then there exists an undominated (at* R*) message* $m_i' \in M_i$ *that weakly dominates* m_i *at* R.

In other words, mechanisms that exploit infinite chains of dominated strategies, as occurs in tailchasing constructions, are ruled out. Note that, like the best response criterion, it is not just a property of the mechanism, but a joint restriction on the mechanism *and the domain.* Jackson[22] (1992) shows that a weaker equilibrium notion than Nash equilibrium, called "undominated strategies" has a similar property to undominated Nash implementation, namely that essentially all social choice correspondences are implementable. He shows that if mechanisms are required to be bounded, then very restrictive results reminiscent of the Gibbard-Satterthwaite theorem hold, so that almost no social choice function is implementable via undominated strategies with bounded mechanisms. However, these negative results do not carry over to undominated Nash implementation.

Following the work of Jackson (1992), Jackson, Palfrey, and Srivastava (1994) provide a characterization of undominated Nash implementation using bounded mechanisms and requiring the best response property. They find that the boundedness restriction, while ruling out some social choice correspondences, is actually quite permissive. First of all, social choice correspondences that are Nash implementable are implementable by bounded mechanisms (see also Tatamatami (1991) and Yamato (1993)). Second, in economic environments with free disposal, any interior allocation rule is implementable. Furthermore, there are many allocation rules that fail to be subgame perfect Nash implementable that are implementable via undominated Nash equilibrium using bounded mechanisms.

2.3 Virtual Implementation

2.3.1 Virtual Nash implementation

A mechanism virtually implements a social choice function[23] if it can (exactly) implement arbitrarily close approximations of that social choice function. The concept was first introduced by Matsushima (1988). It is immediately obvious that, regardless of the domain and regardless of the equilibrium concept, the set of virtually implementable social choice functions contains the set of all implementable social choice functions. What is less obvious, is how much more is virtually implementable compared with what is exactly implementable. It turns out that it makes a big difference.

[22]That is also the first paper to seriously raise the issue of mixed strategies. All of the results that have been described so far in this paper are for *pure strategy implementation.* Only very recently have results been appearing that explicitly address the mixed strategy problem. See for example the work by Abreu and Matsushima (1992a) on virtual implementation.

[23]The work on virtual implementation limits attention to single valued social choice correspondences. Since the results in this area are so permissive (i.e., few social choice functions fail to be virtually implementable), this does not seem to be an important restriction.

One way to see why so much more is virtually implementable can be seen by referring back to Figure 3. That figure shows how the preferences R and R' must line up in order for monotonicity to have any bite in pure exchange economies. As can readily be seen, this is not a generic picture. Rather, Figure 4 shows the generic case, where monotonicity places no restrictions on the social choice at R' if $a = f(R)$. Virtual implementation exploits the nongenericity of situations where monotonicity is binding.[24] It does so by implementing lotteries that produce, in equilibrium at R, $f(R)$ with very high probability, and some other outcomes with very low probability.

In finite or countable economic environments, every social choice function is virtually implementable if individuals have preferences over lotteries that admit a von Neumann-Morgenstern representation and if there are at least three agents.[25] The result is proved in Abreu and Sen (1991) for the case of strict preferences and under a domain restriction that excludes unanimity among the preferences of the agents over pure alternatives. They also address the 2-agent case, where a nonempty lower intersection property is needed.

A key difference between the virtual implementation construction and the Nash implementation construction has to do with the use of lotteries instead of pure alternatives in the test pairs. In particular, virtual implementation allows test pairs involving lotteries in the neighborhood (in lottery space) of $f(R)$ rather than requiring the test pairs to *exactly involve* $f(R)$. It turns out that by expanding the first allocation of the test pair to any neighborhood of $f(R)$, one can always find a test pair of the sort required in the definition of monotonicity.

There are several ways to illustrate why this is so. Perhaps the simplest is to consider the case of von Neumann-Morgenstern preferences for lotteries. If an individual maximizes expected utility, then his indifference surfaces in lottery space are parallel hyperplanes. For the case of three pure alternatives, this is illustrated in Figure 6 below.

[FIGURE 6 HERE]

For this three alternative case, consider two preference profiles, R and R', which differ in some individual's von Neumann-Morgenstern utility function. This means that the slope of the indifference lines for this individual have changed. Accordingly, in *every* neighborhood of *every* interior lottery in Figure 6, there exists a test pair of lotteries such that this agent has a preference switch over the test pair of lotteries. Now consider a social choice function that assigns a pure allocation to each preference profile, but which fails to satisfy monotonicity. In other words, the social choice function assigns one of the vertices of the triangle in Figure 6 to each profile. We can perturb this social choice

[24]This fact that monotonicity holds generically is proved formally in Bergin and Sen (1992). They show for classical pure exchange environments with continuous, strictly monotone (but not necessarily convex) preferences there exists a dense subset of utility functions that always "cross" (i.e., there are never tangencies of the sort depicted in Figure 2).

[25]In fact, more general lottery preferences can be used, as long as they satisfy a condition that guarantees individuals prefer lotteries that place more probability weight on more-preferred pure alternatives.

function ever so slightly so that instead of assigning a vertex, it assigns an interior lottery, x, arbitrarily close to the vertex. This "approximation" of the social choice function satisfies monotonicity because there exists agent i (whose von Neumann-Morgenstern utilities have changed), and a lottery y such that $xR_iyP_i'x$. In this way, every (interior) approximation of every pure social choice function in this simple example is monotonic and hence (if veto power problems are avoided) implementable.

Abreu and Sen (1991) prove that this simple construction outlined above for the case of von Neumann-Morgenstern preferences and $|A| = 3$ is very general. The upshot of this is that moving from exact to virtual implementation completely eliminates the necessity of monotonicity.

2.3.2 Virtual implementation in iterated removal of dominated strategies

An even more powerful result is established in Abreu and Matsushima (1992a). They show that by requiring only virtual implementation, then in finite-profile environments one can find mechanisms such that not only is there a unique Nash equilibrium that approximately yields the social choice function, but the Nash equilibrium is strictly dominance solvable. They exploit the fact that for each agent there is a function from his possible preferences to lottery space, h, such that if $R_i \neq R_i'$, then $h(R_i)P)ih(R_i')$ and $h(R_i')P_i'h(R_i)$. The message space of agent i consists of a single report of i's own preferences and multiple (ordered) reports of the entire preference profile, with the final outcome patching together pieces of a lottery, each piece of which is determined by some portion of the reported profiles. The payoff function is then constructed so that falsely reporting one's own type as R' at R will lead to individual i receiving the $g(R')$ lottery instead of the $g(R)$ lottery with some probability, so this false report is a strictly dominated strategy. Incentives are provided so that subsequent[26] reports of the profile will "agree" with earlier reports in a particular way. The first defection from agreement is punished severely enough to make it unprofitable to defect from the truthful self-reports of the first component of the message space. The degree of approximation can then be made as fine as one wishes simply by requiring a very large number of reported profiles.

Formally, a message for i, is a $K+1$ vector $m_i = (m_i^0, m_i^1, \ldots, m_i^K)$ where the first component is an element of i's set of possible preferences and the other K components are each elements of the set of possible preference profiles. The outcome function is then pieced together in the following way. Let ε be some small positive number. With probability ε/I (where I is the number of players), the outcome is based only on m_i^0, and equals $h(m_i^0)$ so i is strictly better off reporting m_i^0 honestly. With probability ε^2/I agent i is rewarded if, for all $k = 1, \ldots, K, m_i^k = m^0$ whenever $m_j^k = m^0$ for all $j \int = i$ for all $h < k$. That is, i gets a small reward (in expected terms) for honestly revealing

[26]The term "subsequent" should not be interpreted as meaning that the profiles are reported sequentially, since the game is simultaneous-move. Rather, the vector of reported profiles is *ordered*, so subsequent refers to reported profile with the next index number. Glazer and Rubinstein (1994) show that there is a similar sequential game that can be constructed which is dominance solvable following similar logic.

his preference, and then gets an order of magnitude *smaller* reward for always agreeing with the vector of first reports (including his own). These are the only pieces of the outcome function that are affected by m^0. Clearly for ε small enough the first order loss overwhelms any possible second order gain from falsely reporting m_i^0. Thus messages involving false reports of m_i^0 are strictly dominated.

The remaining pieces of the outcome function (each of which is used with probability $(1-\varepsilon-\varepsilon^2)/K$) correspond to the final K components of the messages, where each agent is reporting a preference profile. If everyone agrees on the k^{th} profile, then that k^{th} piece of the outcome function is simply the social choice function at that commonly reported profile. For K large enough, the gain one can obtain from deviating and reporting $m_i^k \neq m^0$ in the k^{th} piece can be made arbitrarily small. But the penalty from being the first to report $m_i^k \neq m^0$ is constant with respect to K, so this penalty will exceed any gain from deviating when K is large. Thus deviating $h = k+1$ can be shown to be dominated once all strategies involving deviations at $h < k+1$ have been eliminated. Variations on this "piecewise" approximation technique also appear in Abreu and Matsushima (1990) where the results are extended to incomplete information (see below) and Abreu and Matsushima (1994) where a similar technique is applied to exact implementation via iterated elimination of *weakly* dominated strategies.[27]

This kind of construction is quite a bit different from the usual Maskin-type of construction used elsewhere in the proofs of implementation theorems. It has a number of attractive features, one of which is the avoidance of any mixed strategy equilibria. In other constructions, mixed strategies are usually just ignored. This can be problematic as an example of Jackson (1992) shows that there are some Nash implementable social choice correspondences that are impossible to implement by a finite mechanism without introducing other mixed strategy equilibria. A second feature is that in finite domains one can implement using finite message spaces. While this is also true for Nash implementation when the environment is finite, there are several examples that illustrate the impossibility of finite implementation in other settings. Palfrey and Srivastava (1991a) show that sometimes infinite constructions are needed for undominated Nash implementation, and Dutta and Sen (1994b) show that Bayesian Nash implementation in finite environments can require infinite message spaces.

Glazer and Rosenthal (1992) raise the issue that in spite of the obvious virtues of the implementing mechanism used in the Abreu and Matsushima (1992a) proof, there are other drawbacks. In particular, Glazer and Rosenthal (1992) argue that the kind of game that is implied by the mechanism is precisely the same kind of game that game theorists have argued as being fragile, in the sense that the predictions of Nash equilibrium are not *a priori* plausible. Abreu and Matsushima (1992b) respond that they believe iterated strict dominance is a good solution concept for predictive[28] purposes, especially in the

[27] Glazer and Perry (1992) show that this mechanism can be reconstructed as a multistage mechanism which can be solved by backward induction. Glazer and Rubinstein (1993) propose that this reduces the computational burden on the players.

[28] In implementation theory, it is the predictive value of the solution concept that matters. One can

context of their construction. However, preliminary experimental findings (Sefton and Yavas, 1993) indicate that in some environments the Abreu-Matsushima mechanisms perform poorly.

This is part of an ongoing debate in implementation theory about the "desirability" of mechanisms and/or solution concepts in the constructive existence proofs that are used to establish implementation results. The arguments by critics are based on two premises: 1) equilibrium concepts, or at least the ones that have been explored, do not predict equally well for all mechanisms; and 2) the quality of the existence result is diminished if the construction uses a mechanism that seems unattractive. Both premises suggest interesting avenues of future research.

An initial response to 1) is that these are empirical issues that require serious study, not mere introspection. The obvious implication is that experimental[29] work in game theory will be crucial to generating useful predictive models of behavior in games. This in turn may require a redirection of effort in implementation theory. For example, from the game theory experiments that have been conducted to date, it is clear that limited rationality considerations will need to be incorporated into the equilibrium concepts, as will statistical (as opposed to deterministic) theories of behavior.[30]

Possible responses to 2) are more complicated. The cheap response is that the implementing mechanisms used in the proofs are not meant to be representative of mechanisms that would actually be used in "real" situations that have a lot of structure. These are merely mathematical techniques, and any mechanism used in a "real" situation should exploit the special structure of the situation. Since the class of environments to which the theorems apply is usually very broad, the implementing mechanisms used in the constructive proofs must work for almost any imaginable setting. The question this response begs is: *for a specific problem of interest, can a "reasonable" mechanism be found?* The existence theorems do not answer this question, nor are they intended to. That is a question of specific application. So far, even with the alternative mechanisms of Abreu-Matsushima, the mechanisms used in general constructive existence theorems are impractical. However, some nice results for familiar environments exist (e.g., Crawford 1979, Moulin 1984, Jackson and Moulin 1990) that suggest we can be optimistic about finding practical mechanisms for implementation in some common economic settings.

---

think of the solution concept as the planner's model for predicting outcomes that will arise under different mechanism and in different environments. If the model predicts inaccurately, then a mechanism will fail to implement the planner's targeted social choice function.

[29]The use of controlled experimentation in settling these empirical questions is urged in Abreu and Matsushima's (1992b) response to Glazer and Rosenthal (1992).

[30]See, for example, McKelvey and Palfrey (1993).

3 Implementation with Incomplete Information

This section looks at the extension of the results of section 3 to the case of incomplete information. Just as most of the results above are organized around Nash equilibrium and refinements, the same is done in this section, except the baseline equilibrium concept is Harsanyi's (1967-68) Bayesian equilibrium.[31]

3.1 The Type Representation

The main difference in the model structure with incomplete information is that a domain specifies not only the set of possible preference profiles, but also the information each agent has about the preference profile and about the other agents' information. We adopt the "type representation" that is familiar to literature on Bayesian mechanism design (see, e.g., Myerson 1985).

An *incomplete information domain*[32] consists of a set, I, of n agents, a set, A, of feasible alternatives, a set of types, T_i, for each agent $i \in I$, a von Neumann Morgenstern utility function for each agent, $u_i : T \times A \to \mathcal{R}$, and a collection of conditional probability distributions $\{q_i(t_{-i}|t_i)$, for each $i \in I$ and for each $t_i \in T_i$. There are a variety of familiar domain restrictions that will be referred to, when necessary, as follows:

1. *Finite types:*	$\|T_i\| < \infty$
2. *Diffuse priors:*	$q_i(t_{-i}\|t_i) > 0$, for all $i \in I$, for all $t_i \in T_i$, and for all $t_{-i} \in T_{-i}$
3. *Private values:*[33]	$u_i(t_i, t_{-i}, a) = u_i(t_i, t'_{-i}, a)$ for all $i, t_i, t_{-i}, t'_{-i}, a$
4. *Independent types:*	$q_i(t_{-i}\|t_i) = q_i(t_{-i}\|t'_i)$ for all $i, t_i, t'_{-i}, t_{-i}, a$
5. *Value-distinguished types:*	For all $i, t_i, t'_i \in T_i, t_i \neq t'_i, \exists a, b$ such that $u_i(t_i, t_{-i}, a) > u_i(t_i, t_{-i}, b)$ and $u_i(t'_i, t_{-i}, b) > u_i(t'_i, t_{-i}, a)$ for all $t_{-i} \in T_{-i}$.

A *social choice function* (or *allocation rule*) $f : TA$ assigns a unique outcome to each type profile. A *social choice correspondence*, F, is a collection of social choice functions. The set of all allocation rules in the domain is denoted by X, so in general, we have $f \in F \subseteq X$. A mechanism $\mu =< M, g >$ is defined as before. A *strategy* for i is a function mapping T_i into M_i, denoted $\sigma_i : T_i \to M_i$. We also denote type t_i of player i's interim utility of an allocation rule $x \in X$ by:

$$U_i(x, t_i) = E_t\{u_i(x(t), t)|t_i\}$$

[31]This should come as no surprise to the reader, since Bayesian equilibrium is simply a version of Nash equilibrium, adapted to deal with asymmetries of information.

[32]Myerson (1985) calls this a Bayesian Collective Decision Problem.

[33]In this case, we simply write $u_i(t_i, a)$, since i's utility depends only on his own type.

where E_t is the expectation over t. Similarly, given a strategy profile σ in a mechanism μ, we denote type t_i of player i's *interim utility of strategy* σ *in* μ by:

$$U_i(\sigma, t_i) = E_t\{u_i(g(\sigma(t)), t)|t_i\}$$

3.2 Bayesian Nash Implementation

Bayesian Nash implementation, like Nash implementation has two components, *incentive compatibility* and *uniqueness*. The main difference is that incentive compatibility imposes genuine restrictions on social choice functions, unlike the case of complete information. When players have private information, the planner must provide the individual with incentives to reveal that information, in contrast to the complete information case, where an individual's report of his information could be checked against another individual's report of that information. Thus, while the constructions with complete information rely heavily on mutual auditing schemes that we called "agreement mechanisms," the constructions with incomplete information do not.[34]

Definition 19 *A strategy* σ *is a* Bayesian equilibrium *of* μ *if, for all* i *and for all* $t_i \in T_i$

$$U_i(\sigma, t_i) \geq U_i(\sigma_i', \sigma_{-i}, t_i) \text{ for all } \sigma_i' : T_i \to M_i$$

Definition 20 *A social choice function* $f : T \to A$ *(or allocation rule* $x : T \to A$*) is* Bayesian implementable *if there is a mechanism* $\mu =< M, g >$ *such that there exists a Bayesian equilibrium of* μ *and, for every Bayesian equilibrium,* σ*, of* $M, f(t) = g(\sigma(t))$ *for all* $t \in T$.

Implementable social choice sets are defined analogously.

For the rest of this section, we restrict attention to the simpler case of *diffuse types*, defined above. Later in the chapter, the extension of these results to more general information structures will be explained.

Incentive Compatibility and the Bayesian Revelation Principle

Paralleling the definition for complete information, a social choice function (or allocation rule) is called (Bayesian) incentive compatible if and only if it can arise as a

[34]There are special exceptions where mutual auditing schemes can be used, which include domains in which there is enough redundancy of information in the group so that an individual's report of the state may be checked against the joint report of the other individuals. This requires a condition called *Non-Exclusive Information* (NEI). See Postlewaite and Schmeidler, 1986 or Palfrey and Srivastava, 1986. Complete information is the extreme form of NEI.

Bayesian equilibrium of some mechanism. The *revelation principle* (Myerson 1979, Harris and Townsend 1981) is the simple proposition that an allocation rule x can arise as the Bayesian equilibrium to some mechanism if and only if truth is a Bayesian equilibrium of the direct[35] mechanism, $\mu =< T, x >$. Thus, we state the following.

Definition 21 *An allocation rule x is* incentive compatible *if, for all i and for all $t_i, t_i' \in T_i$*

$$U_i(x, t_i) \geq E_t\{u_i(x(t_i', t_{-i}), t)|t_i\}$$

Uniqueness

Just as the multiple equilibrium problem can arise with complete information, the same can happen with incomplete information. In particular, direct mechanisms often have this problem (as was the case with the "agreement" direct mechanism in the complete information case). Consider the following example.

Example 22 (an allocation rule investigated in Holmström and Myerson (1983))

There are two agents, each of whom has two types. Types are equally likely and statistically independent and individuals have private values. The alternative set is $A = \{a, b, c\}$. Utility functions are given by (u_{ij} denotes the utility to type j of player i):

$$u_{11}(a) = 2u_{11}(b) = 1u_{11}(c) = 0 \quad u_{12}(a) = 0u_{12}(b) = 4u_{12}(c) = 9$$

$$u_{21}(a) = 2u_{21}(b) = 1u_{21}(c) = 0 \quad u_{22}(a) = 2u_{22}(b) = 1u_{22}(c) = -8$$

The following social choice function, f, is incentive compatible and efficient (where f_{ij} denotes the outcome when player 1 is type i and player 2 is type j):

$$f_{11} = a \quad f_{12} = b \quad f_{21} = c \quad f_{22} = b.$$

It is easy to check that for the direct revelation mechanism $< T, f >$, there is a "truthful" Bayesian equilibrium where both players adopt strategies of reporting their actual type, i.e., f is incentive compatible. However, there is another equilibrium of $< T, f >$, where both players always report type 2 and the outcome is always b. We call such strategies in the direct mechanism *deceptions*, since such strategies involve falsely reported types. Denoting this deceptive strategy profile as α, it defines a new social choice function which we call f_α defined by $f_\alpha(t)f(\alpha(t))t$. This illustrates that this particular allocation rule is not Bayesian Nash implementable by the direct mechanism. However,

[35] A mechanism is *direct* if $M_i = T_I$ for all $i \in I$.

it turns out to be possible to add messages, augmenting[36] the direct mechanism into an "indirect" mechanism that implements f. One way to do this is by giving player 1 another pair of messages, call them "truth" and "lie" one of which must be sent along with the report of his type. The outcome function is then defined so that $g(m) = f(t)$ if the vector of reported types is t and player one says "truth." If player 1 says "lie," then $g(m) = f(t_1, t_2')$ where t_1 is player 1's reported type and t_2' is the opposite of player 2's reported type. This is illustrated in Figure 7 below.

[FIGURE 7 HERE]

It is easy to check that if the players use the α deception above, then player 1 will announce "lie," which is not an equilibrium since player 2 would be better off always responding by announcing type 1. In fact, simple inspection shows that there are no longer any Bayesian equilibria that lead to social choice functions different from f, and (truth, "truth") is a Bayesian equilibrium[37] that leads to f.

Given that the incentive compatibility condition holds, the implementation problem boils down to determining for which social choice functions it is possible to augment the direct mechanism as in the example above, to eliminate unwanted Bayesian equilibria. This is the so-called *method of selective elimination* (Mookherjee and Reichelstein 1990) that is used in most of the constructive sufficiency proofs in implementation theory. Again paralleling the complete information case, there is a simple necessary condition for this to be possible, which is an "interim" version of Maskin's monotonicity condition (definition 6), called *Bayesian monotonicity.* △

Definition 23 *A social choice correspondence F is* Bayesian monotonic *if, for every $f \in F$ and for every joint deception $\alpha : T \to T$ such that $f_\alpha \notin F, \exists i \in I, t_i \in T_i$, and an allocation rule $y : T \to A$ such that $U_i(f_\alpha, t_i) < U_i(y_\alpha, t_i)$ and, for all $t_i' \in T_i, U_i(f, t_i') \geq U_i(y, t_i')$.*

The intuition behind this condition is simpler than it looks. In particular, think of the relationship between f and f_α being roughly the same in the above definition as the relationship between R and R', the difference being that with asymmetric information, we need to consider changes in the entire social choice function f, rather than limiting attention to the particular change in type profile from R to R' (or t to t', in the type notation). So, if $f_\alpha \notin F$ (analogous to $a \notin F(R')$ in the complete information formulation), we need a test agent, i, and a test allocation *rule* y (analogous to test allocation, in monotonicity definition), such that i's (interim) preference between f and y is the reverse of his preference between f_α and y_α (with the appropriate quantifiers and qualifiers included). Thus the basic idea is the same, and involves a test agent and a test allocation rule.

[36] The terminology "augmented" mechanism is due to Mookherjee and Reichelstein (1990).

[37] There is another Bayesian equilibrium that also leads to f. See Palfrey and Srivastava (1993).

3.3 Necessary Condition for Bayesian Implementation

We are now ready to state the main result regarding necessary conditions[38] for Bayesian implementation.

Theorem 24 *If F is Bayesian Nash implementable, then F is Bayesian monotonic, and every $f \in F$ is incentive compatible.*

Proof: The necessity of incentive compatibility is obvious. The proof for necessity of Bayesian monotonicity follows the same logic as the proof for necessity of monotonicity with complete information (see Theorem 8). ∎

3.4 Sufficiency Theorems

As with complete information, sufficient conditions generally require an allocation rule to be in the social choice correspondence if there is nearly unanimous agreement among the individuals about the "best" allocation rule. This is the role of NVP in the original Maskin sufficiency theorem. There are two ways to guarantee this. The first way (and by far the simplest) is to make a domain assumption that avoids the problem by ruling out preference profiles where there is nearly unanimous agreement. The prototypical example of such a domain is a pure exchange economy. In that case, there is a great deal of conflict across agents, as each agent's most preferred outcome is to be allocated the entire societal endowment, and this most preferred outcome is the same regardless of the agent's type. Another related example includes the class of environments with sidepayments using a divisible private good that everyone values positively, the best known case being quasi-linear utility. We present two sufficiency results, one for the case of pure exchange economies, and the second for a generalization. We assume throughout that information is diffuse and $n \geq 3$.

Consider a pure exchange economy with asymmetric information, E, with L goods and n individuals, where the societal endowment[39] is given by $w = (w_1, \ldots, w_L)$. The alternative set, A, is the set of all nonnegative allocations of w across the n agents.[40] The set of feasible allocation rules mapping T into A are denoted X.

[38]There are other necessary conditions. For example, F must be closed with respect to common knowledge concatenations. See Postlewaite and Schmeidler (1986) or Palfrey and Srivastava (1993) for details.

[39]We will not be addressing questions of individual rationality, so the initial allocation of the endowment is left unspecified.

[40]One could permit free disposal as well, but this is not needed for the implementation result. The constructions by Postlewaite and Schmeidler (1986) and Palfrey and Srivastava (1989a) assume free disposal. We do not assume it here, but do assume diffuse information. Free disposability simplifies the constructions when information is not diffuse, by permitting destruction of the entire endowment (i.e., all agents receive 0) when the joint reported type profile is not consistent with any type profile in T.

Theorem 25 *Assume $n \geq 3$ and information is diffuse. A social choice function $x \in X$ is Bayesian Nash implementable if and only if it satisfies incentive compatibility and Bayesian monotonicity.*

Proof: Only if follows from Theorem 24. It is only slightly more difficult. Once again, we use a variation on the agreement/integer game, adapted to the incomplete information framework. Notice, however, that there is always "agreement" with diffuse types, since each player is submitting a different component of the type profile, and all reported type profiles are possible. Each player is asked to report a type and either an allocation rule that is *constant* in his own type (but can depend on other players' types) or a nonnegative integer. Thus:

$$M_i = T_i x \{X_{-i} \cup \{0, 1, \ldots\}\}$$

Where X_{-i} denotes the set of allocation rules that are constant with respect to t_i. The *agreement region* is the set of message profiles where each player sends a reported type and "0." The *unilateral disagreement region* is the set of message profiles where exactly one agent reports a type and something other than "0." Finally, the *disagreement region* is the set of all message profiles with at least two agents failing to report "0." In the agreement region, the outcome is just $x(t)$, where t is the reported type profile. In the unilateral disagreement region the outcome is also just $x(t)$, unless the disagreeing agent, i, sends $y \in X_{-i}$ with the property that $U_i(x, t_i') \geq U_i(y, t_i')$ for all $t_i' \in T_i$. In that case, the outcome is $y(t)$. In the disagreement region, the agent who submits the highest integer[41] is allocated w and everyone else is allocated 0.

Notice how the mechanism parallels very closely the complete information mechanism. The structure of the unilateral disagreement region is such that if all individuals are reporting truthfully, no player can unilaterally falsely report and or disagree and be better off. By incentive compatibility it does not pay to announce a false type. The fact that y does not depend on the disagreer's types implies that it doesn't pay to report y and a false (or true) type. Therefore, there is a Bayesian equilibrium in the agreement region, where all players truthfully report their types. There can be no equilibrium outside the agreement region, because there would be at least two agents each of whom could unilaterally change their message and receive w. Thus the only possible other equilibria that might arise would be in the agreement region, where agents are using a joint deception α. But the Bayesian monotonicity condition (which f satisfies by assumption) says that either $x_\alpha = x$ or there exists a y, and i, and a t_i, such that $U_i(x, t_i') \geq U_i(y, t_i')$ for all $t_i' \in T_i$ but $U_i(y_\alpha, t_i) > U_i(x_\alpha, t_i)$. Since it is easy to project y onto X_{-i} (see Palfrey and Srivastava 1993) and preserve these inequalities, it follows that i is better off deviating unilaterally and reporting y instead of "0." ∎

[41] In this region, if a player sends an allocation instead of an integer, this is counted as "0." Ties are broken in favor of the agent with the lowest index.

The extension of the above result to more general environments is simple, as long as individuals have different "best elements" that do not depend on their type. For each i, suppose that there exists an alternative b_i such that $U_i(b_i, t) \geq U_i(a, t)$ for all $a \in A$ and $t \in T$, and further suppose that for all i, j it is the case that $U_i(b_i, t) > U_i(b_j, t)$ for all $t \in T$. If this condition holds, we say players have distinct best elements.

Theorem 26 *If $n \geq 3$, information is diffuse and players have distinct best elements, then f is Bayesian implementable if and only if f is incentive compatible and Bayesian monotonic.*

Proof: Identical to the proof of Theorem 25, except in the disagreement region the outcome is b_i, where i is winner of the integer game. ∎

Jackson (1991) and Matsushima (1990) show that the condition of distinct best elements can be further weakened, and their result is summarized in Palfrey and Srivastava (1993, p. 35). An even more general version, that considers nondiffuse as well as diffuse information structures is in Jackson (1991). That paper identifies a condition that is a hybrid between Bayesian monotonicity and an interim version of NVP, called *monotonicity-no-veto* (MNV). The earlier papers by Postlewaite and Schmeidler (1986) and Palfrey and Srivastava (1987, 1989a) also consider nondiffuse information structures.

Dutta and Sen (1991b) provide a sufficient condition for Bayesian implementation, when $n \geq 3$ and information is diffuse, that is even weaker than the MNV condition of Jackson (1991). They call this condition *extended unanimity*, and it, like MNV, incorporates Bayesian monotonicity. They also prove that when this condition holds and T is finite, then any incentive compatible social choice function can be implemented using a finite mechanism. They do this using a variation on the integer game, called a *modulo* game,[42] which accomplishes the same thing as an integer game but only requires using the first n positive integers.

Dutta and Sen (1994b) raise an interesting point about the size of the message space that may be required for implementation of a social choice function. They present an example of a social choice function that fails their sufficiency condition (it violates unanimity and there are only two agents), but is nonetheless implementable via Bayesian equilibrium. But they are able to show that the only implementing mechanisms use infinite message spaces, in spite of the fact that both A and T are finite.

Dutta and Sen (1994a) extend their general characterization of Bayesian implementable social choice correspondences when $n \geq 3$ to the $n = 2$ case, using an interim version of the nonempty lower intersection property that they used in their $n = 2$ characterization with complete information (Dutta and Sen, 1991a). This complements some earlier work on characterizing implementable social choice functions by Mookherjee and Reichelstein (1990). Dutta and Sen (1994a) extend these results to characterize

[42]The modulo game is due to Saijo (1988) and is also used in McKelvey (1989) and elsewhere. A potential weakness of a modulo game is that it typically introduces unwanted mixed strategy equilibria that could be avoided by the familiar greatest-integer game.

Bayesian implementable social choice correspondences for the $n = 2$ case, for "economic environments."[43]

All of the results described above are restricted (either explicitly or implicitly) to finite sets of player types. Obviously for many applications in economics this is a strong requirement. Duggan (1994b) provides a rigorous treatment of the many difficult technical problems that can arise when the space of types is uncountable. He extends the results of Jackson (1991) to very general environments, and identifies some new, more inclusive conditions that replace previous assumptions about best elements, private values, and economic environments. The key assumption he uses is called *interiority*, which is satisfied in most applications.

3.5 Implementation using refinements of Bayesian equilibrium

Just as in the case of complete information, refinements permit a wider class of social choice functions to be implemented. These fall into two classes: dominance based refinements using simultaneous-move mechanisms, and sequential rationality refinements using sequential mechanisms. In both cases, the results and proof techniques have similarities to the complete information case.

3.5.1 Undominated Bayesian equilibrium

The results for implementation using dominance refinements in the normal form are limited to undominated Bayesian implementation, where a nearly complete characterization is given in Palfrey and Srivastava (1989b). An undominated Bayesian equilibrium is a Bayesian equilibrium where no player is using a weakly dominated strategy. There are several results, some positive and some negative. First, in private value environments with diffuse types and value-distinguished types, any incentive compatible allocation rule satisfying no veto power is implementable via undominated Bayesian equilibrium. The proof assumes the existence of best and worst elements[44] for each type of each agent, but does not require No Veto Power. They also show that with non-private values, some additional very strong restrictions are needed, and, moreover, the assumption of value distinction is critical.[45]

[43]The term *economic* is vague. "Informally speaking, an economic environment is one where it is possible to make some individual strictly better off from any given allocation in a manner which is independent of her type. This hypothesis, while strong, will be satisfied if there is a transferable private good in which the utilities of both individuals are strictly increasing." (Dutta and Sen, 1994a, p. 52.

[44]Notice that if A is finite or more generally if A is compact and preferences are continuous, then best and worst elements exist. The proof can be extended to cover some special environments where best elements do not exist, such as the quasi-linear utility case.

[45]The assumption of value distinction is stronger than might appear. It rules out environments where two types of an agent differ only in their beliefs about the other agents. One can imagine some natural environments where value distinction might be violated, such as financial trading environments, where a key feature of the information structure involves what agents know about what other agents know.

Two simple voting public goods examples illustrate both the power and the limitations (with common values) of the undominated refinement.

Example 27 There are three agents, two feasible outcomes, $A = \{a, b\}$, private values, independent types, and each player can be one of two types. Type α strictly prefers a to b and type β strictly prefers b to a, and the probability of being type α is q, with $q^2 >= \frac{1}{2}$. The "best solution" according to almost any set of reasonable normative criteria is to chose a if and only if at least two agents are type α. Surprisingly, this "majoritarian" solution, while incentive compatible, is not implementable via Bayesian equilibrium. It is fairly easy to show that for any mechanism that produces the majoritarian solution as a Bayesian equilibrium that mechanism will have another equilibrium in which outcome b is produced at every type profile. However, it is easy to see that the majoritarian solution is implementable via *undominated* Bayesian equilibrium, since it is the unique undominated Bayesian equilibrium[46] outcome in the direct mechanism. $\triangle$

Example 28 This is the same as Example 27 (two feasible outcomes, three agents, two independent types and type α occurs with probability $q^2 > \frac{1}{2}$.), except there are common values.[47] The common preferences are such that if a majority of agents are type α, then everyone prefers a to b, and if a majority of agents are type β, then everyone prefers b to a. We call these "majoritarian preferences." Obviously, there is a unique best social choice function for essentially any non-malevolent welfare criterion, which is the majoritarian (and *unanimous*, as well!!) solution: choose a if and only if at least two agents are type α.

First observe that because of the common values feature of this example, players no longer have a dominant strategy in the direct game for agents to honestly report their true type. (Of course, truth is still a Bayesian equilibrium of the direct game.) One can show (Palfrey and Srivastava 1989b) that this social choice function is not even implementable in undominated Bayesian equilibrium. In particular, any mechanism which produces the majoritarian solution as an undominated Bayesian equilibrium always has another undominated Bayesian equilibrium where the outcome is always b. $\triangle$

The point of Example 28 is to illustrate that with common values, using refinements may have only limited usefulness in a Bayesian framework. We know from the work in complete information that implementation requires the existence of test agents and test pairs of allocations that involve (often delicate) patterns of preference reversal between preference profiles. Analogously, in the Bayesian setting such preference reversals must occur across *type* profiles. With private values, such preference reversals are easy to find. With common values and/or non-value-distinguished types, such preference reversals often simply do not exist, even in very natural examples of social choice functions that

[46]Notice that it is actually a *dominant strategy equilibrium* of the direct mechanism. This example illustrates how it is possible for an allocation rule to be dominant strategy implementable (and strategy proof), but not Bayesian Nash implementable.

[47]By common values we mean that every agent has the same type-contingent preferences. A related mechanism design problem is explored in more depth by Glazer and Rubinstein (1994).

satisfy No Veto Power and $N > 2$. We turn next[48] to the question of implementation using sequential rationality refinements, where results parallel (to an extent) the results for subgame perfect implementation.

For virtual implementation, Abreu and Matsushima have an extension of their complete information paper on the use of iterated elimination of strictly dominated strategies (Abreu and Matsushima 1990) for implementation in incomplete information environments. They show that, under a condition they call *measurability* and some additional minor restrictions on the domain, any incentive compatible social choice function defined on finite domains can be virtually implemented by iterated elimination of strictly dominated strategies. They conjecture in Abreu and Matsushima (1994) that with some additional assumptions (such as the ability to use small monetary transfers) one can obtain exact implementation via iterated elimination of weakly dominated strategies in finite incomplete information domains.

Duggan (1994a, 1995a) looks at the related issue of virtual implementation in Bayesian equilibrium (rather than iterated elimination of dominated strategies). He shows that the measurability of Abreu and Matsushima (1990) is not necessary for virtual Bayesian implementation. With a mild domain restriction,[49] Bayesian incentive compatibility is necessary and sufficient in environments where there exists some uniquely[50] Bayesian incentive compatible allocation rule. An analogous result is established for dominant strategy implementation, using a similar proof technique.

3.5.2 Implementation via sequential equilibrium

There are two papers that partially characterize the set of implementable social choice functions for incomplete information environments using the equilibrium refinement of sequential equilibrium. The main idea behind these characterizations is the same as the ideas behind the results for subgame perfect equilibrium implementation under conditions of complete information. Instead of requiring a test pair involving the social choice function, x, as is required in Bayesian monotonicity, all that is needed is *some* (interim) preference reversal between *some* pair of allocation rules, plus an appropriate sequence of allocation rules that indirectly connect x with the test pair of allocation rules.

The details of the conditions analogous to indirect monotonicity for incomplete information are messy to state, because of quantifiers and qualifiers that relate to the posterior beliefs an agent could have at different stages of an extensive form game in

[48]There is a second approach to using dominance based refinements in games of incomplete information, which is virtual implementation via iterated elimination of dominated strategies (Abreu and Matsushima 1990). Results parallel their findings for complete information, and the differences are discussed in the Palfrey and Srivastava (1993) monograph. Duggan (1994a) has extended those results to allow for continuous types.

[49]The restriction on the environment is that there must exist some Bayesian incentive compatible allocation rule whose associated direct mechanism has a unique equilibrium.

[50]That is, truth is the unique equilibrium of the allocation rule's direct mechanism.

which different players are adopting different deceptions. However, the intuition behind the condition is similar to the intuition behind Condition α in Abreu and Sen (1990).

As with the necessary and sufficient results for Bayesian implementation, results are easiest to state and prove for the special case of economic environments, where No Veto Power problems are assumed away and where there are at least three agents.

Bergin and Sen (1993) have some results for this case. They identify a condition which is sufficient for implementation by sequential equilibrium in a two-stage game. That paper also makes the point that with incomplete information there exist social choice functions that are implementable sequentially, but are not implementable via undominated Bayesian equilibrium (or via iterated dominance), and are not even virtually implementable. This contrasts with the complete information case, where that any social choice function in economic environments is implementable via subgame perfect Nash equilibrium is also implementable via undominated Nash equilibrium. They are able to obtain these very strong results by showing that the "consistent beliefs" condition of sequential equilibrium can be exploited to place restrictions on equilibrium play in the second stage of the mechanism.

Baliga (1993) also looks at implementation via sequential equilibrium, limited to finite stage extensive games. His paper makes additional restrictions of private values and independent types, which lead to a significant simplification of the analysis.

A more general approach is taken in Brusco (1993) who does not limit himself to stage games nor to economic environments. He looks at implementation via perfect Bayesian equilibrium and obtains the incomplete information equivalent to indirect monotonicity which he calls "Condition $\beta+$." This condition is then combined with No Veto Power in a manner similar to Jackson's (1991) monotonicity-no-veto condition to produce *sequential monotonicity-no-veto.* His main theorem[51] is that any incentive compatible social choice function satisfying SMNV is implementable in perfect Bayesian equilibrium. He also identifies a weaker condition than $\beta+$ (called condition β), which he proves is necessary for implementation in perfect Bayesian equilibrium. However, Brusco's (1993) results are weaker than Bergin-Sen (1993) because his conditions on the requisite sequence of test allocations include a universal quantifier on beliefs that makes it much more difficult to guarantee existence of the sequence. Bergin-Sen show that the very tight condition of belief consistency can replace the universal quantifier. Loosely speaking, Brusco's results exploit *only* the sequential rationality part of sequential equilibrium, while Bergin-Sen exploit both sequential rationality *and* belief consistency. This seemingly minor distinction actually makes quite a difference in proving what can be implemented.

Duggan (1995b) focuses on sequentially rational implementation[52] in quasi-linear en-

[51]The main theorem is stated more generally. In particular he allows for social choice correspondences, which means that the additional restriction of closure under the common knowledge concatenation is required.

[52]Duggan (1995b) defines *sequentially rational implementation* as implementation simultaneously in Perfect Bayesian Equilibrium and Sequential Equilibrium.

vironments where the outcomes are lotteries over a finite set of public projects (and transfers). He shows that any incentive compatible social choice function is implementable in private values environments with diffuse priors over a finite type space if there are three or more agents. He also shows that these results can be extended, with some modifications, in a number of directions: two agents; the domain of exchange economies; infinite type spaces; bounded transfers; nondiffuse priors; and social choice correspondences. He also shows how a "belief revelation mechanism" can be used if the planner does not know the agents prior beliefs, as long as these prior beliefs are common knowledge among the agents.

4 Open Issues

Open problems in implementation theory abound. Several issues have been explored at only a superficial level, and others have not been studied at all. Some of these have been mentioned in passing in the previous sections. There are also numerous untied loose ends having to do with completing full characterizations of implementability under the solution concepts discussed above.

Implementation using perfect information extensive forms and voting trees

Among these uncompleted problems is implementation via backward induction and the closely related problem of implementing using voting trees. If this class of implementation problems has a shortcoming, it is that extensions to the more challenging problem of implementation in incomplete information environments are limited. The structure of the arguments for backward induction implementation fails to extend nicely to incomplete information environments, as we know from the literature on sophisticated voting with incomplete information (e.g., Ordeshook and Palfrey 1988).

Renegotiation and information leakage

Many of the above constructions have the feature that undesirable (e.g., Pareto inefficient, grossly inequitable, or individually irrational) allocations are used in the mechanism to break unwanted equilibria. The simplest examples arise when there exists a universally bad outcome that is reverted too in the event of disagreement. This is not necessarily a problem in some settings, where the planner's objective function may conflict with Pareto optimality from the point of view of the agents (as in many principal-agent problems). However, in some settings, most obviously exchange environments, one usually thinks of the mechanism as being something that the players themselves construct in order to achieve efficient allocations. In this case, one would expect agents to renegotiate outcomes that are commonly known among themselves to be Pareto dominated.[53] Maskin

[53]One doesn't have to look very hard to find counterexamples to this in the real world. Institutional structures (such as courts) are widely used to enforce ex post inefficient allocations in order to provide

and Moore (1989) examine the implications of requiring that the outcomes always to be Pareto optimal. This approach has the virtue of avoiding mixed strategy equilibria and also avoiding implausibly bad outcomes off the the equilibrium path. It has the defect of implicitly permitting the outcome function of the mechanism to depend on the preference profile, which makes the specification of renegotiation somewhat arbitrary. Ideally, one would wish to specify a bargaining game that would arise in the event that an outcome is reached which is inefficient.[54] But then the bargaining game itself should be considered part of the mechanism, which leads to an infinite regress problem.

More generally, the planner may be viewed directly as a player, who has state-contingent preferences over outcomes, just as the other players do. The planner has prior beliefs over the states or preference profiles (even if the players themselves have complete information). Given these priors, the planner has induced preferences over the allocations. This presents a commitment problem for the planner, since the outcome function must adhere to these preferences. This places restrictions both on the mechanism that can be used and also on the social choice functions that can be implemented. Several papers have been written recently on this subject, which vary in the assumptions they make about the extent to which the planner can commit to a mechanism, the extent to which the planner may update his priors after observing the reported messages, and the extent to which the planner participates directly in the mechanism. The first paper on this subject is by Chakravorti, Corchon, and Wilkie (1994) and assumes that the social choice function must be consistent with some prior the planner might have over the states, which implies that the outcome function is restricted to the range of the social choice function. The planner is not an active participant and does not update beliefs based on the messages of the players, nor does he choose an outcome that is optimal given his beliefs. Baliga, Corchon, and Sjöström (1995) obtain results with an actively participating planner,[55] who acts optimally, given the messages of the players and cannot commit ex ante to an outcome function. Thus, the mechanism consists of a message space for the players and a planner's strategy of how to assign messages to outcomes. This strategy replaces the familiar outcome function, but is required to be sequentially rational. Thus, the mechanism is really a two-stage (signalling) game, and an equilibrium of the mechanism must satisfy the the conditions of Perfect Bayesian Equilibrium. Baliga and Sjöström (1995) obtain further results on interactive implementation with an uninformed planner who can commit to an outcome function, and who also participates in the message-sending stage.

Another sort of renegotiation arises in incomplete information settings if the social choice function calls for allocations that are known by at least some of the players to be inefficient. In particular, for certain type realizations, some players may be able to propose to replace the mechanism with a new one that all other types of all other players would unanimously prefer to the outcome of the social choice function. This

salient incentives. Some forms of criminal punishments, such as incarceration and physical abuse, fall into this category.

[54]See, for example, Aghion, Dewatripont, and Rey (1994) or Rubinstein and Wolinsky (1991).

[55]That is, the planner also submits messages. They call this "interactive implementation."

problem of lack of *durability*[56] (Hölmstrom and Myerson 1983) opens up another kind of renegotiation problem, which may involve the potential leakage of information between agents in the renegotiation process. This has been addressed to some extent in principal-agent settings by Maskin and Tirole (1992) and in exchange economies by Palfrey and Srivastava (1993).

Also related to this kind of renegotiation is the problem of preplay communication among the agents. It is well known that preplay communication can expand the set of equilibria of a game, and a similar thing can happen in mechanism design. This can occur because information can be transmitted by preplay communication and because communication opens up possibilities for coordination that were impossible to achieve with independent actions. In nearly all of the implementation theory research, it is assumed that preplay communication is impossible (i.e., the message space of the mechanism specifies all possible communication). An exception is Palfrey and Srivastava (1991b), which explicitly looks at designing mechanisms that are "communication proof," in the sense that the equilibria with arbitrary kinds of preplay communication are interim-payoff-equivalent to the equilibria without preplay communication. They show that for a wide range of economic environments one can construct communication proof mechanisms to implement any interim efficient, incentive compatible allocation rule.

Implementation in dynamic environments

Many mechanism design problems and allocation problems involve intertemporal allocations. One obvious example is bargaining when delay is costly. In that case both the split of the pie and the time of agreement are economically important components of the final allocation. Recently, Rubinstein and Wolinsky (1991) look at the renegotiation proof problem in implementation theory by appending an infinite horizon bargaining game with discounting to the end of each inefficient terminal node. This is an alternative approach to the same renegotiation problem that Maskin and Moore (1989) were concerned about. However, like the rest of implementation theory, their interest is in implementing static allocation rules (i.e., no delay) in environments that are (except for the final bargaining stages) static. This is true for all the other sequential game constructions in implementation theory: time stands still while the mechanism is played out.

Intertemporal implementation raises additional issues. Consider, for example, a setting in which every day the same set of agents is confronted with the next in a series of connected allocation problems, and there is discounting. A preference profile is not an infinite sequence of "one shot" profiles corresponding with each time period. A social choice function is a mapping from the set of these profile sequences into allocation sequences. Renegotiation proofness would impose a natural time consistency constraint that the social choice function would have to satisfy from time t onward, for each t. With this kind of structure one could begin to look at a broader set of economic issues related to growth, savings, intertemporal consumption, and so forth.

[56]Closely related to this are the notions of *ratifiability* and *secure allocations* (Cramton and Palfrey 1994) and *stable allocations* (Legros 1990).

There are some very simple intertemporal allocation problems that could be investigated as a first step. One example is the one-sector growth model of Boylan et al. (1990) which compares different political mechanisms for deciding on investments. As a second example, Bliss and Nalebuff (1984) look at an intertemporal public goods problem. There is a single indivisible public good which can be produced once-and-for-all at any date $t = 1, 2, 3 \ldots$, and preferences are quasilinear with discounting. The production technology requires a unit of private good for the public good to be provided. Thus, an allocation is a time at which the public good is produced and an infinite stream of taxes for each individual, as a function of the profile of preferences for the public good. Bliss and Nalebuff (1984) look at the equilibrium of a specific mechanism, the voluntary contribution mechanism. At each point in time an individual must decide whether or not to privately pay for the public good, depending on their type. The unique equilibrium is for types that prefer the public good more strongly to pay earlier. Thus the public good is always produced by having the individual with the strongest preference for the public good paying for it, and the time of delay before production depends on what the highest valuation is and on the distribution of types. One could generalize this as a dynamic implementation problem, which would raise some interesting questions: What other allocation rules are implementable in this setting? Is the Bliss-Nalebuff (1984) equilibrium allocation rule interim incentive efficient?[57]

Robustness of the mechanism

Implementation theory (even the relaxed problem of virtual implementation) so far has investigated special deterministic models of individual behavior. The key assumption for obtaining results is that the equilibrium model that is assumed to govern individual behavior under any mechanism *is exactly correct.* Many of the mechanisms have no room for error. One would generally think of such fragile mechanisms as being nonrobust. Similarly (especially in the Bayesian environments) the details of the environment, such as the common knowledge priors of the players and the distribution of types, are known to the planner *precisely.* Often mechanisms rely on this exact knowledge. It should be the case that if the model of behavior or the model of the environment is not completely accurate, the equilibrium behavior of the agents does not lead to outcomes too far from the social choice function one is trying to implement.

This problem suggests a need to investigate mechanisms that either do not make special use of detailed information about the environment (such as the distribution of types) or else look at models that permit statistical deviation from the behavior that is predicted under the equilibrium model. In the latter case, it may be more natural to think of social choice functions as type-contingent random variables rather than as *deterministic* functions of the type profile. Related to the problem of robustness of the mechanisms and the possible use of statistical notions of equilibrium is bounded rationality. The usual rationale for purely rational modelling in economics is that it is a good first cut on the problem and may often capture much of the reality of a given

[57]Notice that it is not ex post efficient since there is always delay in producing the public good.

economic situation. In any case, most economists regard the fully rational equilibrium as an appropriate benchmark in most situations. Unfortunately, since implementation theorists and mechanism designers get to *choose* the economic game in very special ways, this rationale loses much of its punch. It may well be that the games where rational models predict poorly are precisely those games that implementation theorist are prone to designing. Integer games, modulo games, "grand lottery" games like those used in virtual implementation proofs, and the enormous message spaces endemic to all the general constructions would seem for the most part to be games that would challenge the limits of even the most brilliant and experienced game player. If such constructions are unavoidable we really need to start to look beyond models of perfectly rational behavior. Even if such constructions are avoidable, we have to be asking more questions about the match (or mismatch) between equilibrium concepts as predictive tools and limitations on the rationality of the players.

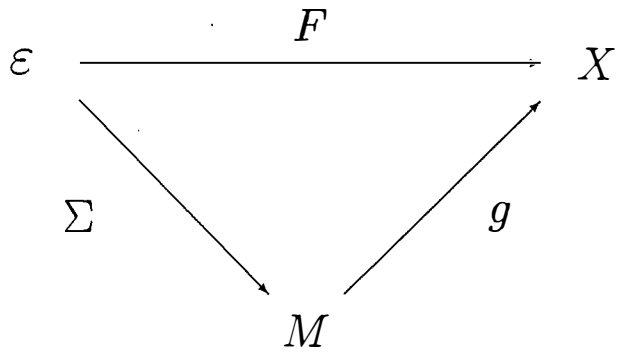

Figure 1. Mount-Reiter Diagram

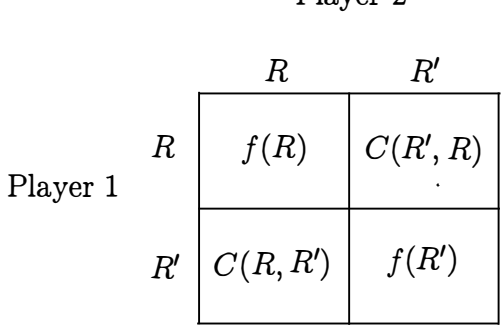

Figure 2. (Weak) Nonempty Lower Intersection:
$f(R)R'C(R,R')\ ,\ f(R')R_2'C(R,R')$
and
$f(R')R_1'C(R',R)\ ,\ f(R)R_2C(R',R)$
$\Rightarrow f(R) \in \boldsymbol{NE}(R)\ ,\ f(R') \in \boldsymbol{NE}(R')$

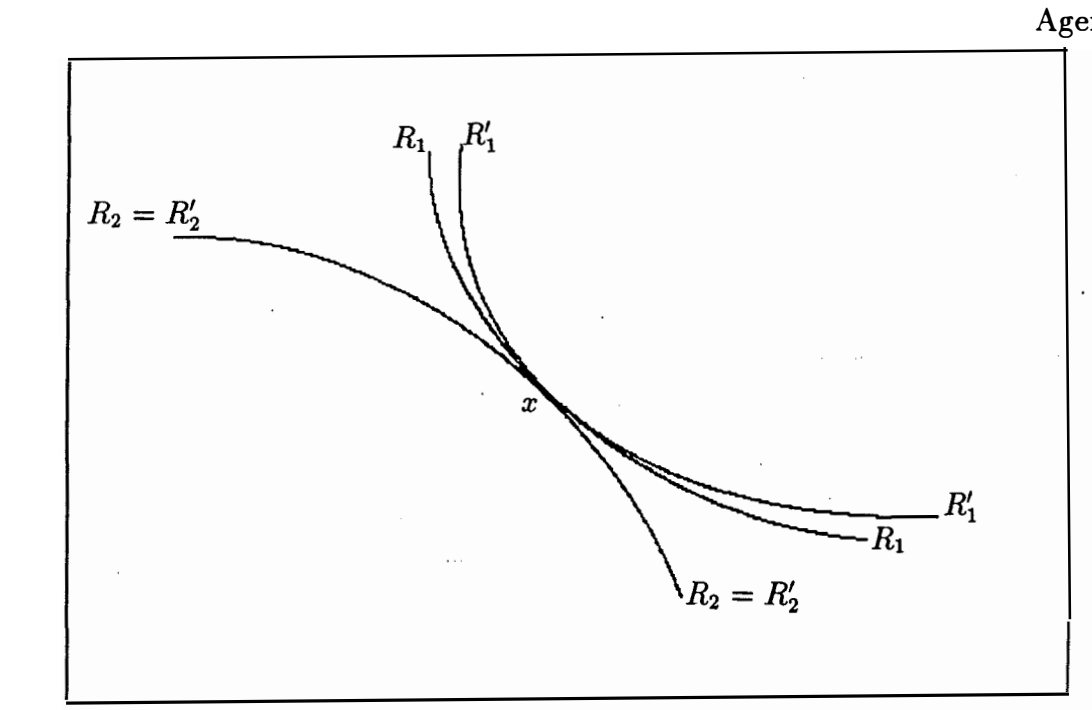

Figure 3. Illustration of monotonicity:
$x \in F(R) \Rightarrow x \in F(R')$

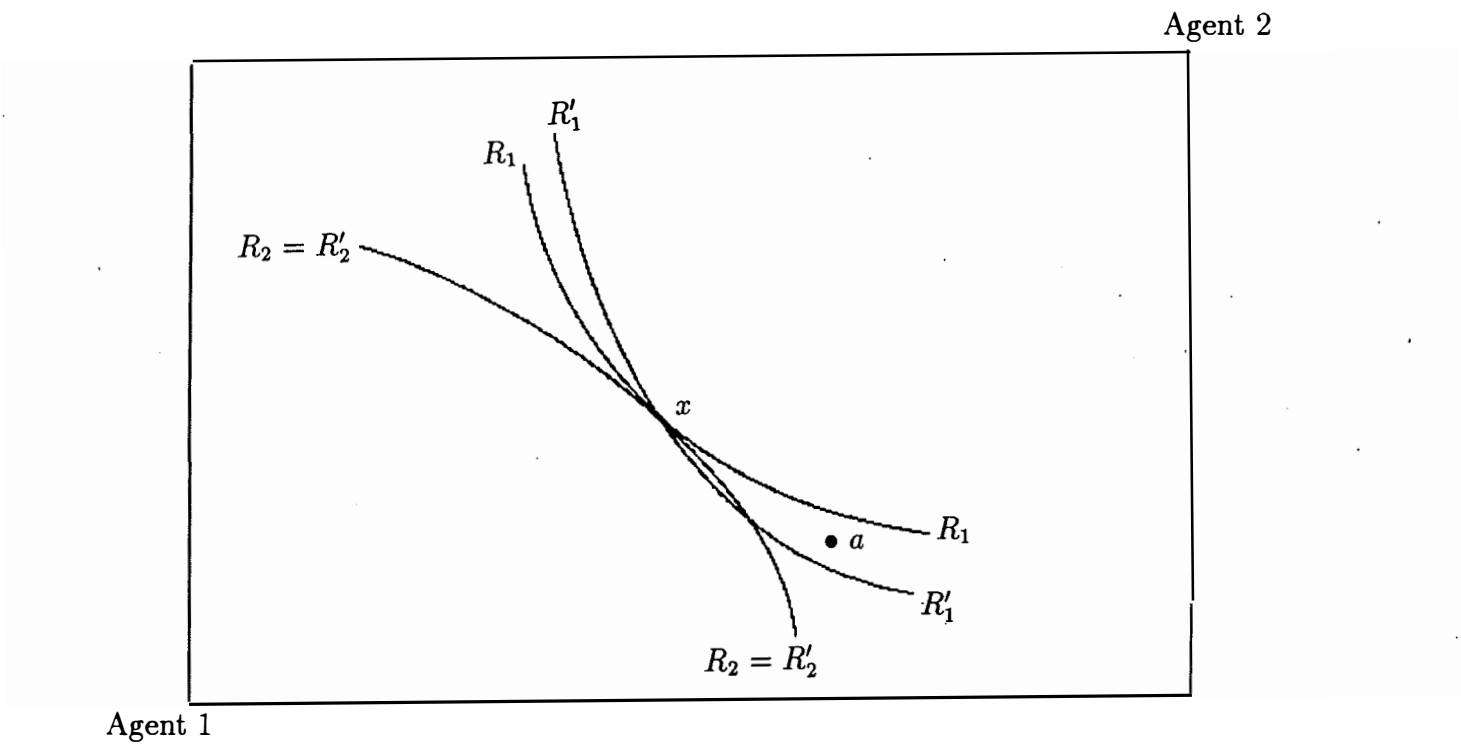

Figure 4. Illustration of test agent and test allocation: $xR_1aP_1'x$

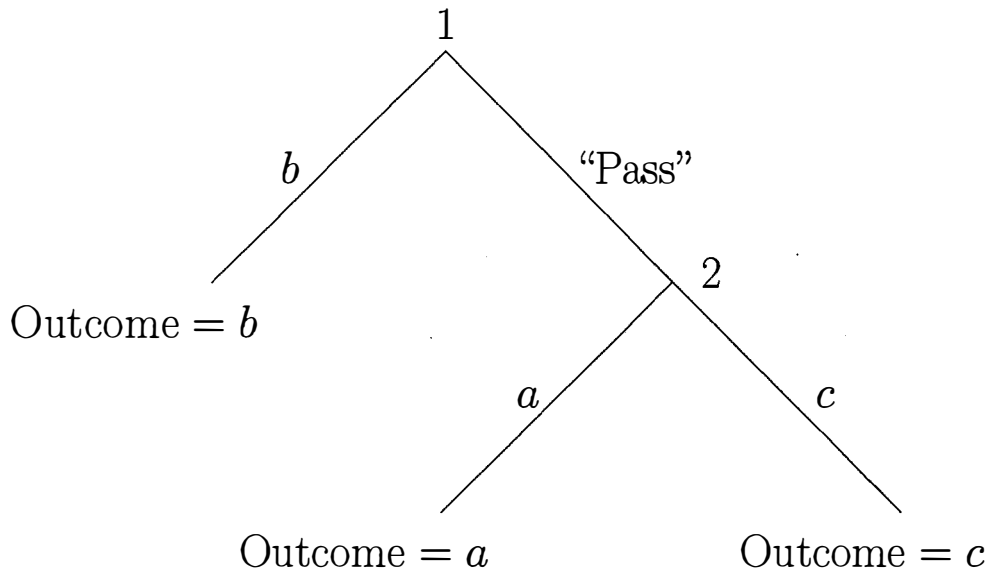

Figure 5. Game tree for example 2.

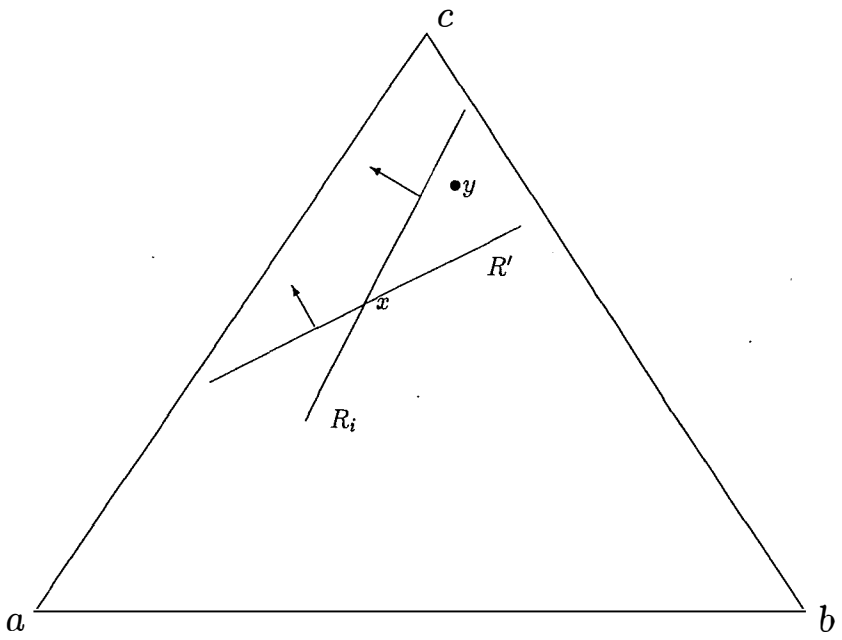

Figure 6. Vertices a, b, and c represent pure alternatives, with all other points representing lotteries over those alternatives. The indifference curves passing through lottery x for agent i under two Von Neumann-Morgenstern utility functions are lablled R_i and R_i', with the direction of preference marked with arrows. Lottery y satisfies $xR_iyP_i'x$.

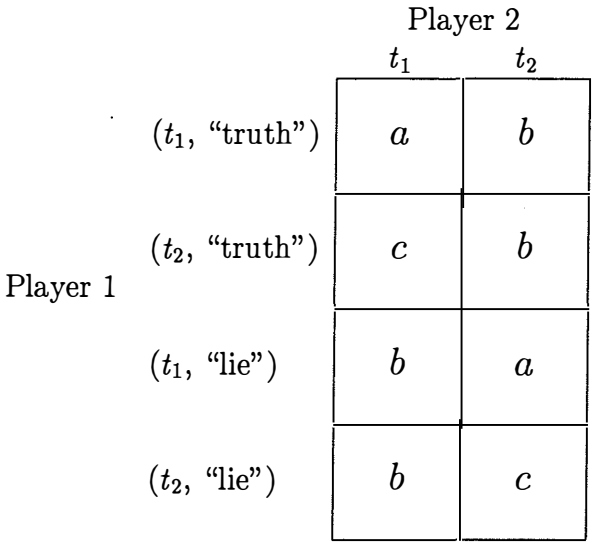

Figure 7. Implementing mechanism for Holmström-Myerson example

References

[1] Abreu, D. and H. Matsushima (1990), "Virtual Implementation in Iteratively Undominated Strategies: Incomplete Information," mimeo.

[2] _____ (1992a), "Virtual Implementation in Iteratively Undominated Strategies: Complete Information," *Econometrica*, 60:993-1008.

[3] _____ (1992b), "A Response to Glazer and Rosenthal," *Econometrica*, 60:1439-42.

[4] _____ (1994), "Exact Implementation," *Journal of Economic Theory*, 64:1–20.

[5] Abreu, D. and A. Sen (1990), "Subgame Perfect Implementation: A Necessary and Almost Sufficient Condition," *Journal of Economic Theory*, 50:285–99.

[6] _____ (1991), "Virtual Implementation in Nash Equilibrium," *Econometrica*, 59:997–1021.

[7] Aghion, P, M. Dewatripont, and P. Rey (1994), "Renegotiation Design with Unverifiable Information," *Econometrica*, 62:257–82.

[8] Baliga, S. (1993), "Implementation in Incomplete Information Environments: The Use of Extension Forms," Discussion Paper, Havard University.

[9] Baliga, S., L. Corchon, and T. Sjöström (1995), "The theory of Implementation when the Planner is a Player," mimeo, Havard University.

[10] Baliga, S. and T. Sjöström (1995), "Interactive Implementation," mimeo, Havard University.

[11] Banks, J. (1985), "Sophisticated Voting Outcomes and Agenda Control," *Social Choices and Welfare*, 1:295–306.

[12] Banks, J., C. Camerer, and D. Porter (1994), "An Experimental Analysis of Nash Refinements in Signalling Games," *Games and Economic Behavior*, 6:1–31.

[13] Bergin, J. and A. Sen (1992), "Implementation in Generic Environments" mimeo.

[14] _____ (1993), "Extensive Form Implementation in Incomplete Information Environments" mimeo.

[15] Bernheim, D., R. Peleg, and M. Whinston (1987), "Coalition-proof Nash Equilibrium, II: Applications," *Journal of Economic Theory*, 42:13–29.

[16] Bliss, C. and B. Nalebuff (1984), "Dragon-Slaying and Ballroom Dancing: The Private Supply of a Public Good," *Journal of Public Economics*, 25:1–12.

[17] Boylan, R., J. Ledyard, and R. McKelvey (1990), "Political Competition in a Model of Economic Growth: Some Theoritical Results," California Institute of Technology, Social Science Working Paper 780.

[18] Brusco, S. (1993), "Perfect Bayesian Implementation" mimeo.

[19] Chakravorti, B., L. Corchon, and S. Wilkie (1994), "Credible Implementation," mimeo.

[20] Chander, P. and L. Wilde (1992) "A General Characterization of Optimal Income Taxation and Enforcement," California Institute of Technology Working Paper 791.

[21] Cramton, P. and T. Palfrey (1994), "Ratifiable Mechanisms: Learning from Disagreement," *Games and Economic Behavior*, in press.

[22] Crawford, V. (1977), "A Game of Fair Division," *Review of Economic Studies*, 44:235–47.

[23] _____ (1979), "A Procedure for Generating Pareto-Efficient Egalitarian Equivalent Allocations," *Econometrica*, 47:49–60.

[24] Danilov, V. (1992), "Implementation via Nash Equilibrium," *Econometrica*, 60:43–56.

[25] Dasgupta, P., P. Hammond, and E. Maskin (1979), "The Implementation of Social Choice Rules: Some General Results on Incentive Compatability," *Review of Economic Studies*, 46:185–216.

[26] Demange, G. (1984), "Implementing Efficient Egalitarian Equivalent Allocations," *Econometrica*, 52:1167–77.

[27] Duggan, J. (1993), "Bayesian Implementation with Infinite Types" mimeo.

[28] _____ (1994a), "Virtual Implementation in Bayesian Equilibrium with Infinite Sets of Types," mimeo, California Institute of Technology.

[29] _____ (1994b), *Bayesian Implementation*, Ph.D. Disertation, California Institute of Technology.

[30] _____ (1995a), "Virtual Bayesian Implementation," mimeo, Queens University.

[31] _____ (1995b), "Sequentially Rational Implementation with Incomplete Information," mimeo, Queens University.

[32] Dutta, B. and A. Sen (1991a), "A Necessary and Sufficient Condition for Two-Person Nash Implementation," *Review of Economic Studies*, 58: 121-8.

[33] _____ (1991b), "Further Results on Bayesian Implementation" mimeo.

[34] _____ (1991c), "Implementation under Strong Equilibrium: A Complete Characterization," *Journal of Mathematical Economics*, 20:49–67.

[35] _____ (1993), "Implementing generalized Condorcet social choice functions via backward induction," *Social Choice and Welfare*, 10:149–60.

[36] _____ (1994a), "Two-Person Bayesian Implementation," *Economic Design*, 1:41–54.

[37] _____ (1994b), "Bayesian Implementation: The Necessity of Infinite Mechanisms," *Journal of Economic Theory*, 64:130–41.

[38] Dutta, B., A. Sen, and R. Vohra (1994), "Nash Implementation Through Elementary Mechanisms in Exchange Economies," *Economic Design*, 1:173–203.

[39] Farquharson, R. (1957/1969), *Theory of Voting*, New Haven: Yale University Press.

[40] Gibbard, A. (1973), "Manipulation of Voting Schemes," *Econometrica*, 41:587–601.

[41] Glazer, J. and C-T. Ma (1989), "Efficient Allocation of a 'Prize' — King Solomon's Dilemma," *Games and Economic Behavior*, 1:222–33.

[42] Glazer, J. and M. Perry (1992), "Virtual Implementation by Backwards Induction," mimeo, Tel Aviv University.

[43] Glazer, J. and R. Rosenthal (1992), "A Note on the Abreu-Matsushima Mechanism," *Econometrica*, 60:1435–8.

[44] Glazer, J. and A. Rubinstein (1993), "Simplicity of Solution Concepts: Subgame Perfect Equilibrium in Extensive Games *vs.* Iteratively Undominated Strategies in Normal Games," mineo, Tel Aviv University.

[45] _____ (1994), "The Design of Organizations for Collecting Information from Conformist Agents," memeo, Tel Aviv University.

[46] Harris, M. and R. Townsend (1981), "Resource Allocation with Asymmetric Information," *Econometrica*, 49:33–64.

[47] Harsanyi, J. (1967, 1968), "Games with Incomplete Information Played by Bayesian Players," *Management Science*, 14:159–82, 320–34, 486–502.

[48] Hererro, M. and S. Srivastava (1992), "Implementation via Backward Induction," *Journal of Economic Theory*, 56:70–88.

[49] Hölmstrom, B. and R. Myerson (1983), "Efficient and Durable Decision Rules with Incomplete Information," *Econometrica*, 51:1799–1819.

[50] Hong, L. (1994), "Bayesian Implementation in Exchange Economies with State Dependent Feasible Sets," mimeo.

[51] Hong, L. and S. Page (1994), "Reducing Informational Costs in Endowment Mechanisms," *Economic Design*, 1:103–17.

[52] Hurwicz, L. (1960), "Optimality and Information Efficiency in Resource Allocation Processes," in K. Arrow, S. Karlin, and P. Suppes (eds.), *Mathematical Methods in the Social Sciences*, Stanford: Stanford University Press, pp. 27–46.

[53] _____ (1977) "Optimality and Informational Efficiency in Resource Allocation Problems," in *Mathematical Methods in the Social Sciences*. (Arrow, Karlin and Suppes, eds.), pp. 27–48, Stanford University Press: Stanford.

[54] _____ (1986), "Incentive Aspects of Decentralization," in K. Arrow and M. Intriligator (eds.), *Handbook of Mathematical Economics*, vol. 3, Amsterdam: North-Holland, pp. 1441–82.

[55] Hurwicz, L., E. Maskin, and A. Postlewaite (1980), "Feasible Implementation of Social Choice Correspondences by Nash Equilibria," mimeo, University of Minnesota.

[56] Jackson, M. (1991), "Bayesian Implementation," *Econometrica*, 59:461–77.

[57] _____ (1992), "Implementation in Undominated Strategies: A Look at Bounded Mechanisms," *Review of Economic Studies*, 59:757–75.

[58] Jackson, M. and H. Moulin (1990), "Implementing a Public Project and Distributing Its Cost," *Journal of Economic Theory*, 57:124–40.

[59] Jackson, M., T. Palfrey, and S. Srivastava (1994), "Undominated Nash Implementation in Bounded Mechanisms," *Games and Economic Behavior*, 6:474–501.

[60] Legros, P. (1990), "Strongly Durable Allocations," CAE Working Paper No. 90-05, Cornell University.

[61] Maskin, E. (1977), "Nash Implementation and Welfare Optimality," mimeo, Massachusetts Institute of Technology.

[62] Maskin, E. and J. Moore (1989), "Implementation with Renegotiation," mimeo.

[63] Maskin, E. and J. Tirole (1992), "The Principal-Agent Relationship with an Informed Principal, II: Common Values," *Econometrica*, 60:1–42.

[64] Matsushima, H. (1988), "A New Approach to the Implementation Problem," *Journal of Economic Theory*, 45:128–44.

[65] _____ (1990), "Characterization of Full Bayesian Implementation," manuscript, Stanford University.

[66] McKelvey, R. (1989), "Game Forms for Nash Implementation of General Social Choice Correspondences," *Social Choice and Welfare*, 6:139–56.

[67] McKelvey, R. and R. Niemi (1978), "A Multistage Game Representation of Sophisticated Voting for Binary Procedures," *Journal of Economic Theory*, 81:1–22.

[68] McKelvey, R. and T. Palfrey (1993), "Quantal Response Equilibria for Normal Form Games," forthcoming in *Games and Economic Behavior.*

[69] Miller, N. (1977), "Graph-Theoretic Approaches to the Theory of Voting," *American Journal of Political Science*, 21:769–803.

[70] Mookherjee, D. and S. Reichelstein (1990), "Implementation Via Augmented Revelation Mechanisms," *Review of Economic Studies*, 57:453–75.

[71] 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.

[72] Moore, J. and R. Repullo (1988), "Subgame Perfect Implementation," *Econometrica*, 46:1191–220.

[73] _____ (1990), "Nash Implementation: A Full Characterization," *Econometrica*, 58:1083-99.

[74] Moulin, H. (1979), "Dominance-Solvable Voting Schemes," *Econometrica*, 47:1337–51.

[75] _____ (1984), "Implementing the Kalai-Smorodinsky Bargaining Solution," *Journal of Economic Theory*, 33:32–45.

[76] _____ (1986) "Choosing From a Tournament," *Social Choice and Welfare*, 3:271–91.

[77] _____ (1993), "Social Choice," in *Handbook of Game Theory*, 2:1091–1125.

[78] Mount, K. and S. Reiter (1974), "The Informational Size of Message Spaces," *Journal of Economic Theory*, 8:161–92.

[79] Myerson, R. (1979), "Incentive Compatibility and the Bargaining Problem," *Econometrica*, 47:61–74.

[80] _____ (1985), "Bayesian Equilibrium and Incentive Compatibility: An Introduction," in L. Hurwicz, D. Schmeidler, and H. Sonnenschein (eds.), *Social Goals and Social Organization: Essays in Memory of Elisha Pazner.* Cambridge: Cambridge University Press, pp. 229–59.

[81] Ordeshook, P. and T. Palfrey (1988), "Agendas, Strategic Voting, and Signaling with Incomplete Information," *American Journal of Political Science*, 32:441–66.

[82] Palfrey, T., and S. Srivastava (1986), "Private Information in Large Economies," *Journal of Economic Theory*, 39:34–58.

[83] ----- (1987), "On Bayesian Implementable Allocations," *Review of Economic Studies*, 54:193–208.

[84] ----- (1989a), "Implementation with Incomplete Information in Exchange Economies," *Econometrica*, 57:115–34.

[85] ----- (1989b), "Mechanism Design with Incomplete Information: A Solution to the Implementation Problem," *Journal of Political Economy*, 97:668–91.

[86] ----- (1991a), "Nash Implemenation Using Undominated Strategies," *Econometrica*, 59:479–501.

[87] ----- (1991b), "Efficient Trading Mechanisms with Preplay Communication," *Journal of Economic Theory*, 55:17–40.

[88] ----- (1993) *Bayesian Implementation*, Harwood Academic Publishers: Reading.

[89] Postlewaite, A. (1985), "Implementation via Nash Equilibria in Economic Environments," in L. Hurwicz, D. Schmeidler and H. Sonnenschein (eds.), *Social Organization: Essays in Memory of Elisha Pazner*, Cambridge: Cambridge University Press, pp. 205–28.

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

[91] Postlewaite, A. and D. Wettstein (1989), "Feasible and Continuous Implementation," *Review of Economic Studies*, 56:603–11.

[92] Reichelstein, S. (1984), "Incentive Compatibility and Informational Requirements," *Journal of Economic Theory*, 34:32–51.

[93] Reichelstein, S. and S. Reiter (1988), "Game Forms with Minimal Strategy Spaces," *Econometrica*, 56:661–92.

[94] Repullo, R. (1987), "A Simple Proof of Maskin's Theorem on Nash Implementation," *Social Choice and Welfare*, 4:39–41.

[95] Rubinstein, A. and A. Wolinsky (1991), "Renegotiation-Proof Implementation and Time Preferences," *American Economic Review*, in press.

[96] Saijo, T. (1988), "Strategy Space Reductions in Maskin's Theorem: Sufficient Conditions for Nash Implementation," *Econometrica*, 56:693–700.

[97] Satterthwaite, M. (1975), "Strategy-proofness and Arrow's Conditions: Existence and Correspondence Theorems for Voting Procedure and Social Welfare Functions," *Journal of Economic Theory*, 10:187–217.

[98] Sefton, M. and A. Yavas (1993), "Abreu-Matsushima Mechanisms: Experimental Evidence," mimeo.

[99] Sen, A (1987), "Two Essays in the Theory of Implementation," Ph.D. Dissertation, Princeton University.

[100] Sjöström, T. (1991a), "Implementation in Undominated Nash Equilibrium without Integer Games," mimeo, Harvard University. To appear in *Games and Economic Behavior.*

[101] _____ (1991b), "On the Necessary and Sufficient Conditions for Nash Implementation," *Social Choice and Welfare*, 8:333–40.

[102] _____ (1993), "Implementation in Perfect Equilibria," *Social Choice and Welfare*, 10:97–106.

[103] Tatamitami (1991), "Double Implementation in Nash and Undominated Nash Equilibrium in Social Choice Environments," mimeo, University of Tsukuba.

[104] Thomson, W. (1979), "Maximin Strategies and Elicitation of Preferences " in J-J. Laffont (ed.), *Aggregation and Revelation of Preferences*, Amsterdam: North-Holland, pp. 245–68.

[105] Tian, G. (1994), "Bayesian Implementation in Exchange Economies with State Dependent Preferences and Feasible Sets," mimeo, Texas A&M University.

[106] Townsend, R. (1979), "Optimal Contracts and Competitive Markets with Costly State Verification," *Journal of Economic Theory*, 21:265–93.

[107] Trick, M. and S. Srivastava (1994), "Sophisticated Voting Rules: The Case of Two Tournaments," Carnegie-Mellon University Working Paper.

[108] Varian, H. (1993), "A Solution to the Problem of Externalities and Public Goods when Agents are Well-Informed," *American Economic Review*, forthcoming.

[109] Wettstein, D. (1992), "Implementation Theory in Economies with Incomplete Information," *Games and Economic Behavior*, 4:463–83.

[110] Williams, S. (1984), "Sufficient Conditions for Nash Implementation," mimeo, University of Minnesota.

[111] Yamato, T. (1993), "Double Implementation in Nash and Undominated Nash Equilibria," *Journal of Economic Theory*, 59:311–23.