# Implementation and Preference for Honesty

Hitoshi Matsushima[1]

Faculty of Economics, University of Tokyo
Hongo, Bunkyo-Ku, Tokyo 113-0033, Japan
e-mail: hitoshi@e.u-tokyo.ac.jp

March 4, 2002
This Version: June 16, 2003

## Abstract

We investigate implementation of social choice functions that map states to lotteries, where agents have preferences not only for consequences but also for 'honesty'. We show that in the complete information environments with three or more agents, every social choice function is implementable in Nash equilibrium. This is in contrast with the standard implementation models where agents have preferences only for consequences and no social choice function depending on factors other than agents' preferences is implementable. We show also that in the incomplete information environments with two or more agents, every Bayesian incentive compatible social choice function can be implemented in Bayesian Nash equilibrium by a mechanism that is universal in the sense that it does not depend on the detail of the private signal structure.

**Keywords**: Preference for Honesty, Normative Social Choice Functions, Implementation, Small Fines, Universal Mechanisms.

[1] The earlier version of this paper corresponds to the old manuscript entitled "Honesty-Proof Implementation". I would like to thank Mr. Daisuke Shimizu for his careful reading. All errors are mine.

# 1. Introduction

We investigate implementation of social choice functions that map states to lotteries over the set of pure alternatives where we use only small fines.[2] Real individuals may have preferences not only for consequences but also for anything non-consequential such as 'honesty'. Based on this fact, this paper will assume that each agent has positive cost for dishonest reporting. Several works such as Erard and Feinstein (1994), Alger and Ma (2003), and Deneckere and Severinov (2001) examined the case that agents' ability to manipulate information is limited, and demonstrated that including agents who have preferences for honesty could significantly alter the model. These works assumed that the cost for dishonest reporting is sufficiently large. In contrast, the present paper will allow the maximal total cost for dishonest reporting to be as close to zero as possible.

First of all, we show that in the complete information environments with three or more agents, *every* social choice function is implementable in Nash equilibrium where honest reporting is always the unique iteratively undominated strategy profile. This positive result is in sharp contrast with the standard implementation models, where every agent has preference only for consequences, [3] as follows. A social choice function is said to be *normative* if it depends on factors other than agents' preferences for consequences. It has been assumed in the standard implementation models with complete information that a social choice function is not normative, and therefore, its domain is restricted to the set of agents' preference profiles. In fact, we can easily check that if every agent has preference only for consequences, then *no* normative social choice function is implementable in Nash equilibrium.[4] This negative result is valid, even if we replace Nash equilibrium with any equilibrium concept with complete information and consider virtual implementation.[5] In contrast with the standard implementation models above, the present paper shows that whenever each agent has preference for honesty in the marginal way, then every normative social choice function is implementable in Nash equilibrium.

Second, we show that in the incomplete information environments with two or more agents, every Bayesian incentive compatible social choice function is implementable in

---

[2] There are excellent surveys on implementation theory such as Moore (1992), Palfrey (1992), and Maskin and Sjostrom (2002).

[3] Eliaz (2001) took into account factors other than individuals' preferences for consequences such as bounded rationality. The paper by Glazer and Rubinstein (1998) is more closely related to the present paper, where agents are allowed to have preference for anything non-consequential.

[4] See Maskin and Sjostrom (2002), for example. Serrano and Vohra (2001) presented an example in the economic environments with incomplete information where agents' preferences are the same across states but their initial endowments depend on the state. In their example, every individually rational social choice function is non-constant, and therefore, is normative in our sense.

[5] Fix any equilibrium concept arbitrarily, where the set of equilibrium outcomes is assumed to be the same between states that have the same preference profiles. Implementability requires the set of equilibrium outcomes to equal the value of the social choice function. If the social choice function is normative, then there exist two distinct states that have the same preference profile, and therefore, have the same set of equilibrium outcomes irrespective of the mechanism construction, but that induces different socially desired lotteries. This is a contradiction of implementability.

Bayesian Nash equilibrium where honest reporting is the unique Bayesian iteratively undominated strategy profile. Of particular importance, the mechanism constructed is regarded as being *universal* in the sense that it does not depend on the detail of the private signal structure. Hence, the central planner can construct the mechanism without knowing what the true private signal structure is. This point is in sharp contrast with the standard implementation models with incomplete information, where a mechanism is tailored to a specific situation, and therefore, is difficult to use in practice.

In the present paper, we will construct mechanisms where each agent makes multiple announcements about the state. In this respect, the paper is related to Abreu and Matsushima (1992a, 1992b). In contrast with Abreu and Matsushima, however, the present paper does not use the device of virtualness originated by Matsushima (1988) and Abreu and Sen (1991). Hence, our results are on 'exact' implementation as opposed to 'virtual', and our mechanisms are much simpler than the mechanisms in Abreu and Matsushima for this reason.[6]

The organization of the paper is as follows. Section 2 shows the basic model. Section 3 investigates the complete information environments and shows that every social choice function is implementable in Nash equilibrium. Section 4 investigates the incomplete information environments and shows that every Bayesian incentive compatible social choice function can be implemented in Bayesian Nash equilibrium by a universal mechanism.

[6] Abreu and Matsushima (1994) studied exact implementation but used an idea similar to virtualness.

## 2. The Basic Model

Let $N = \{1,...,n\}$ denote the finite set of agents where $n \geq 2$. We will denote $n+1=1$ and $1-1=n$. Let $\Omega$ and $A$ denote the nonempty and compact sets of states and alternatives, respectively. Let $M_i$ denote the set of messages for each agent $i \in N$. Let $M = \prod_{i \in N} M_i$. At every state $\omega \in \Omega$, the agents will announce a message profile $m = (m_1,...,m_n) \in M$ and the central planner will decide an alternative $a \in A$ and fine each agent $i \in N$ a nonnegative amount of private goods $-t_i \in [0,\varepsilon]$, where $\varepsilon > 0$ is the upper bound of fines. The resultant utility for each agent $i \in N$ is given by $u_i(a,t_i,m_i,\omega)$, where we allow each agent to have intrinsic value on which message to be announced.

A *simple lottery* is denoted by $\alpha : A \to [0,1]$, which has the countable support $\Gamma(\alpha) \subset A$ where $\sum_{a \in \Gamma(\alpha)} \alpha(a) = 1$. Let $\Delta$ denote the set of all simple lotteries. We assume the expected utility hypothesis where the expected utility for agent $i$ at state $\omega \in \Omega$ when the central planner chooses an alternative according to $\alpha \in \Delta$ and fines agent $i$ the amount $-t_i$ of private goods is denoted by $u_i(\alpha,t_i,m_i,\omega) \equiv \sum_{a \in A} u_i(a,t_i,m_i,\omega)\alpha(a)$. A *social choice function* is defined as a mapping from states to simple lotteries and is denoted by $f : \Omega \to \Delta$, where $f(\omega)$ is regarded as the socially desired lottery at state $\omega$. We denote by $F$ the set of all social choice functions. A *mechanism* is defined by $G = (g,x)$, where $g : M \to \Delta$, $x = (x_i)_{i \in N}$, and $x_i : M \to [-\varepsilon,0]$ for each $i \in N$. When the agents announce a message profile $m \in M$, the central planner will choose an alternative according to the lottery $g(m) \in \Delta$ and fine each agent $i \in N$ the amount $-x_i(m) \in [0,\varepsilon]$.

## 3. Complete Information

This section considers the situation in which all agents have complete information about the state. A message profile $m \in M$ is said to be a *Nash equilibrium* in the game defined by $(G, \omega)$ if for every $i \in N$,

$$u_i(g(m), x_i(m), m_i, \omega) \geq u_i(g(m_i', m_{-i}), x_i(m_i', m_{-i}), m_i', \omega) \text{ for all } m_i' \in M_i .$$

A social choice function $f \in F$ is said to be *implemented by a mechanism* $G = (g, x)$ *in Nash equilibrium* if at every state $\omega \in \Omega$, there exists a Nash equilibrium in $(G, \omega)$, and every Nash equilibrium $m \in M$ in $(G, \omega)$ induces the socially desired lottery, i.e.,

$$g(m) = f(\omega), \text{ and } x_i(m) = 0 \text{ for all } i \in N .$$

We specify

$$M_i = \Omega^K ,$$

where $K$ is a positive integer. Each agent $i \in N$ announces $K$ elements of $\Omega$ at one time. Let $M_i = M_i^1 \times \cdots \times M_i^K$ and $m_i = (m_i^1, ..., m_i^K)$ where $M_i^k = \Omega$ and $m_i^k \in M_i^k$. For every $k \in \{1, ..., K\}$, let $m^k = (m_1^k, ..., m_n^k)$. The *honest message for each agent* $i \in N$ *at each state* $\omega \in \Omega$ is defined as $\mu_i(\omega) = (\mu_i^k(\omega))_{k=1}^K \in M_i$ where $\mu_i^k(\omega) = \omega$ for all $k \in \{1, ..., K\}$. Let $\mu(\omega) = (\mu_i(\omega))_{i \in N}$ denote the honest message profile.

We assume

$$n \geq 3 .$$

We will specify a mechanism $G = G^*(f, K, \varepsilon)$ as follows. Fix a lottery $\overline{\alpha} \in \Delta$ arbitrarily. We define $z : \Omega^n \to \Delta$ in ways that for every $\omega \in \Omega$ and every $\delta = (\delta_1, ..., \delta_n) \in \Omega^n$,

$$z(\delta) = f(\omega) \text{ if } \delta_i = \omega \text{ for at least } n-1 \text{ components of } \delta ,$$

and for every $\delta \in \Omega^n$,

$$z(\delta) = \overline{\alpha} \text{ if there exists no such } \omega \in \Omega .$$

We specify $g$ by

$$g(m) = \frac{\sum_{k=1}^{K-1} z(m^k)}{K-1} \text{ for all } m \in M .$$

For every $k \in \{1, ..., K-1\}$, the central planner will choose an alternative according to the lottery $z(m^k)$ with probability $\frac{1}{K-1}$, where for every $\omega \in \Omega$,

$$z(m^k) = f(\omega) \text{ if } m_i^k = \omega \text{ for at least } n-1 \text{ agents,}$$

and

$$z(m^k) = \overline{\alpha} \text{ if there exists no such } \omega \in \Omega .$$

For every $i \in N$, we specify $x_i$ by

$$x_i(m) = -\varepsilon \text{ if there exists } k \in \{1, ..., K-1\} \text{ such that } m_i^k \neq m_{i-1}^K \text{ and } m_j^h = m_{j-1}^K \text{ for}$$

all $j \in N$ and all $h \in \{1,...,k-1\}$,

and

$x_i(m) = 0$ if there exists no such $k \in \{1,...,K-1\}$.

Each agent $i \in N$ is fined the amount $\varepsilon$ if and only if she is the first to announce a different opinion from agent $(i-1)'s$ K-th opinion. When the agents announce $\mu(\omega)$, the central planner will choose an alternative according to $f(\omega)$ and no agents are fined, i.e., for every $\omega \in \Omega$,

$g(\mu(\omega)) = f(\omega)$, and $x_i(\mu(\omega)) = 0$ for all $i \in N$.

Let $l_i(m_i, \omega) \in \{0,...,K\}$ denote the number of $k \in \{1,...,K\}$ satisfying $m_i^k \neq \omega$, i.e., the number of agent $i's$ dishonest announcements. We specify $u_i(a, t_i, m_i, \omega)$ by

$$u_i(a, t_i, m_i, \omega) = v_i(a, \omega) + t_i - \frac{l_i(m_i, \omega)}{K} c,$$

where $c > 0$ [7]. Note that $v_i(a, \omega)$ represents agent $i's$ preference for consequences, and that each agent has preference for honesty in the sense that the cost of reporting each component of her message dishonestly is given by $\frac{c}{K}$. Here, each agent concerns not only whether she announces dishonestly but also how often she announces dishonestly amongst her K announcements.

**Theorem 1:** *Suppose*

(1) $$\max_{(a,a',\omega,i) \in A^2 \times \Omega \times N} \left| v_i(a, \omega) - v_i(a', \omega) \right| < (K-1)\varepsilon.$$

*Then, every social choice function $f \in F$ is implemented by $G^*(f, K, \varepsilon)$ in Nash equilibrium.*

**Proof:** Fix $\omega \in \Omega$ arbitrarily. Note that each agent $i \in N$ has incentive to announce $m_i^K = \mu_i^K(\omega)$ in the game $(G^*(f, K, \varepsilon), \omega)$, because both $g(m)$ and $x_i(m)$ are independent of $m_i^K$ and she has preference for honesty.[8] Fix $k \in \{1,...,K-1\}$ and $m \in M$ arbitrarily, where

$m_i^{k'} = \mu_i^{k'}(\omega) = \omega$ for all $i \in N$ and all $k' \in \{1,...,k-1,K\}$.

Fix $i \in N$ arbitrarily, and suppose $m_i^k \neq \mu_i^k(\omega)$. Let $m_i' \in M_i$ be the message for agent $i$ defined by

$m_i'^k = \mu_i^k(\omega)$,

and

$m_i'^{k'} = m_i^{k'}$ for all $k' \in \{1,...,K\} / \{k\}$.

If $m_j^k = \mu_j^k(\omega)$ for all $j \in N / \{i\}$, then, it follows that $g(m)$ is independent of $m_i^k$ and

---

[7] We can extend our results to the general case where the cost for dishonest reporting depends on $(i, \omega) \in N \times \Omega$, i.e., we can replace $c$ with $c_i(\omega) > 0$ with no substantial changes.

[8] When agents have no preferences for honesty, we have to use another device such as virtualness, in order to incentivize them to make their K-th announcements honestly, which makes implementation of normative social choice functions impossible.

$x_i(m_i', m_{-i}) \geq x_i(m)$, which, together with the fact that agent $i$ has preference for honesty, implies that agent $i$ has incentive to announce $m_i'$ instead of $m_i$. If $m_j^k \neq \mu_j^k(\omega)$ for some $j \neq i$, then it follows that $x_i(m_i', m_{-i}) - x_i(m) = \varepsilon$, which, together with the inequality (1), implies that agent $i$ has incentive to announce $m_i'$ instead of $m_i$, because

$$\begin{aligned} & u_i(g(m), t_i(m), m_i, \omega) - u_i(g(m_i', m_{-i}), t_i(m_i', m_{-i}), m_i', \omega) \\ &= \frac{1}{K-1}\{v_i(z(m^k), \omega) - v_i(z(\mu_i^k(\omega), m_{-i}^k), \omega)\} - \varepsilon \\ &\leq \frac{1}{K-1} \max_{(a,a',\omega,i) \in A^2 \times \Omega \times N} |v_i(a, \omega) - v_i(a', \omega)| - \varepsilon < 0. \end{aligned}$$

The above arguments imply that for every $\omega \in \Omega$, $\mu(\omega)$ is the unique iteratively undominated message profile in $(G^*(f, K, \varepsilon), \omega)$. Hence, we have proved that every social choice function $f \in F$ is implemented by $G^*(f, K, \varepsilon)$ in Nash equilibrium.

**Q.E.D.**

We must note that for every $\varepsilon > 0$, there exist a positive integer $K$ that satisfies the inequality (1). Hence, it follows that *for every $\varepsilon > 0$ and every $c > 0$, there exists a positive integer $K$ such that every social choice function $f \in F$ is implemented by the mechanism $G^*(f, K, \varepsilon)$ in Nash equilibrium.*[9]

The proof of Theorem 1 shows that for every $\omega \in \Omega$, the honest message profile $\mu(\omega)$ is the unique iteratively undominated strategy profile in the game $(G^*(f, K, \varepsilon), \omega)$. This implies that $f$ is implementable not only in mixed Nash equilibrium but also in iteratively undominated strategies.

We must note that the construction of the mechanism $G^*(f, K, \varepsilon)$, together with the inequality (1), does not depend on the cost $c > 0$. This implies that, *irrespective of how close to zero the cost $c > 0$ is, the mechanism $G^*(f, K, \varepsilon)$ can implement $f$ in iteratively undominated strategies*.

A social choice function $f \in F$ is said to be *normative* if it depends on factors other than agents' preferences for honesty in the sense that there exist $\omega \in \Omega$ and $\omega' \in \Omega / \{\omega\}$ such that

$$f(\omega) \neq f(\omega'),$$

and for every $i \in N$,

$$v_i(\cdot, \omega) = b v_i(\cdot, \omega') + d \text{ for some } b > 0 \text{ and some } d \in R.$$

We must note that if every agent has preference only for consequences, i.e., $c = 0$, then no

[9] Deneckere and Severinov (2001) constructed mechanisms in which each agent makes multiple announcements. In their paper, the cost for each agent's making all announcements dishonestly becomes as large as possible as the number of her announcements increases. In contrast, the present paper assumes that the total cost of each agent's making all announcements dishonestly is set constant (maybe close to zero) irrespective of how many announcements each agent will make.

normative social choice function is implementable in Nash equilibrium. This negative result is valid, even if we replace Nash equilibrium with any equilibrium concept with complete information and consider virtual implementation. In contrast, whenever $c > 0$, then we need no such restrictions, and therefore, every normative social choice function is implementable in iteratively undominated strategies.

Besides utilitarian interpersonal utility comparisons, there exist many previous attempts to establish various ideas on theoretical foundation of social choice and welfare that are based on factors other than individuals' preferences, such as primary goods originated by John Rawls (1971), liberty, functioning, and capabilities originated by Amartya Sen (1982, 1985, 1999), compensation and responsibility originated by Ronald Dworkin (1981), and others. With no doubt it will be a most substantial ongoing and future research in the social choice theory literature to cultivate further the influence of factors other than individuals' preferences on the social welfare judgments.[10] In order to make such non-preference based social choice theories in practice, however, individuals must have incentive to announce their reports honestly on these non-preference factors as well as their preferences. The present paper is regarded as a first attempt to establish affirmative answers to this manipulation problem.

We do not need to require every relevant agent to participate in the decision procedure described by the mechanism $G^*(f, K, \varepsilon)$. All we need to require is that at least three agents who have complete information about the state participate in the decision procedure. Hence, we can assume $n = 3$, even if there exist other individuals who can not participate in the decision procedure but whose preferences have influence on which lottery to be socially desired. This point is in contrast with the standard implementation models where all relevant individuals must participate in the decision procedure.

[10] For example, see Basu, Pattanaik, and Suzumura (1995), Sen (1999), Suzumura (2002), and others.

## 4. Incomplete Information

This section considers the incomplete information environments. Each agent $i \in N$ receives her private signal $\omega_i \in \Omega_i$ where $\Omega_i$ is the nonempty and finite set of private signals. The set of states is defined as the Cartesian product of the sets of private signals, i.e., $\Omega \equiv \prod_{i \in N} \Omega_i$. This section assumes $n \geq 2$.[11]

A *message rule for each agent* $i \in N$ is defined as a function $\eta_i : \Omega_i \to M_i$. Let $\Xi_i$ denote the set of all message rules for agent $i$. We denote by $\eta = (\eta_i)_{i \in N}$ a message rule profile. Let $\Xi \equiv \prod_{i \in N} \Xi_i$, $\eta(\omega) = (\eta_i(\omega_i))_{i \in N}$, and $\eta_{-i}(\omega_{-i}) = (\eta_j(\omega_j))_{j \in N/\{i\}}$. A *private signal structure* is defined by $p = (p_i(\cdot \mid \omega_i))_{i \in N, \omega_i \in \Omega_i}$ where $p_i(\cdot \mid \omega_i) : \Omega_{-i} \to [0,1]$ is the conditional probability function. Let $P$ denote the set of private signal structures. A message rule profile $\eta \in \Xi$ is said to be a *Bayesian Nash equilibrium in a mechanism* $G$ *for a private signal structure* $p \in P$ if for every $i \in N$ and every $\omega_i \in \Omega_i$,

$$E[u_i(g(\eta(\omega)), x_i(\eta(\omega)), \eta_i(\omega_i), \omega) \mid p, \omega_i]$$
$$\geq E[u_i(g(m_i, \eta_{-i}(\omega_{-i})), x(m_i, \eta_{-i}(\omega_{-i})), m_i, \omega) \mid p, \omega_i] \text{ for all } m_i \in M_i,$$

where $E[\cdot \mid p, \omega_i]$ implies the expected value conditional on $p$ and $\omega_i$. A social choice function $f \in F$ is said to be implemented by a mechanism $G$ in Bayesian Nash equilibrium for a private signal structure $p \in P$ if there exists a Bayesian Nash equilibrium in $G$ for $p$, and every Bayesian Nash equilibrium $\eta$ in $G$ for $p$ always induces the socially desired lottery, i.e., for every $\omega \in \Omega$,

$$g(\eta(\omega)) = f(\omega), \text{ and } x_i(\eta(\omega)) = 0 \text{ for all } i \in N.$$

We specify

$$M_i = \Omega_i^K \text{ for all } i \in N.$$

Each agent $i \in N$ announces $K$ elements of $\Omega_i$ at one time. Let $M_i = M_i^1 \times \cdots \times M_i^K$ where $M_i^k = \Omega_i$. The *honest message rule for each agent* $i \in N$ is defined by $\hat{\eta}_i = (\hat{\eta}_i^k)_{k=1}^K \in \Xi_i$ where $\hat{\eta}_i^k(\omega_i) = \omega_i$ for all $k \in \{1,...,K\}$ and all $\omega_i \in \Omega_i$. Let $\hat{\eta} = (\hat{\eta}_i)_{i \in N}$ denote the honest message rule profile.

Fix a positive integer $\hat{K} \in \{1,...,K-1\}$ and a positive real number $\hat{\varepsilon} \in (0, \varepsilon]$ arbitrarily. We will specify a mechanism $G = \hat{G}(f, K, \hat{K}, \hat{\varepsilon}, c)$ as follows. For every $m \in M$,

$$g(m) = \frac{\sum_{k=1}^{\hat{K}} f(m^k)}{\hat{K}}.$$

[11] This section assumes $n \geq 3$.

For every $k \in \{1,...,\hat{K}\}$, the central planner will choose an alternative according to the lottery $f(m^k)$ with probability $\frac{1}{\hat{K}}$. For every $i \in N$, and $m \in M$,

$x_i(m) = -\hat{\varepsilon}$ if there exist $k \in \{1,...,\hat{K}\}$ and $k' \in \{\hat{K}+1,...,K\}$ such that $m_i^k \neq m_i^{k'}$, and $m_j^h = m_j^{h'}$ for all $j \in N$, all $h \in \{1,...,k-1\}$, and all $h' \in \{\hat{K}+1,...,K\}$,

and

$x_i(m) = 0$ if there exists no such $(k,k')$.

Each agent $i \in N$ is fined the amount $\hat{\varepsilon}$ if and only if she is the first to announce a contradictory opinion to what she announces as her last $K - \hat{K}$ opinions. When the agents announce $\hat{\eta}(\omega)$, the central planner will choose an alternative according to $f(\omega)$ and no agents are fined, i.e., for every $\omega \in \Omega$,

$g(\hat{\eta}(\omega)) = f(\omega)$, and $x_i(\hat{\eta}(\omega)) = 0$ for all $i \in N$.

Let $q_i(m_i, \omega_i) \in \{0,...,K\}$ denote the number of $k \in \{1,...,K\}$ satisfying $m_i^k \neq \omega_i$. In a similar way to Section 3, we specify $u_i(a, t_i, m_i, \omega)$ by

$$u_i(a, t_i, m_i, \omega) = v_i(a, \omega) + t_i - \frac{q_i(m_i, \omega_i)}{K} c .$$

**Theorem 2:** *Suppose that*

(2) $$(\frac{K - \hat{K}}{K})c > \hat{\varepsilon},$$

*and*

(3) $$\max_{(a,a',\omega,i) \in A^2 \times \Omega \times N} \left| v_i(a, \omega) - v_i(a', \omega) \right| < \frac{\hat{K}\hat{\varepsilon}}{2}.$$

*Then, for every $p \in P$, every social choice function $f \in F$ is implemented by $\hat{G}(f, K, \hat{K}, \hat{\varepsilon}, c)$ in Bayesian Nash equilibrium if for every $i \in N$ and every $\omega_i \in \Omega_i$,*

(4) $$E[v_i(f(\omega), \omega) \mid p, \omega_i] > E[v_i(f(\omega_i', \omega_{-i}), \omega) \mid p, \omega_i] - \frac{\hat{K}}{K} c \text{ for all } \omega_i' \in \Omega_i .$$

**Proof:** Fix $\eta \in \Xi$, $i \in N$, and $\omega_i \in \Omega_i$ arbitrarily, where $\eta_i^k(\omega_i) \neq \hat{\eta}_i^k(\omega_i)$ for some $k \in \{\hat{K}+1,...,K\}$. Let $m_i \in M_i$ denote the message for agent $i$ defined by

$m_i^k = \hat{\eta}_i^k(\omega_i)$ for all $k \in \{\hat{K}+1,...,K\}$,

and

$m_i^k = \eta_i^k(\omega_i)$ for all $k \in \{1,...,\hat{K}\}$.

Suppose that $\eta_i^k(\omega_i) \neq \eta_i^{\hat{K}+1}(\omega_i)$ for some $k \in \{\hat{K}+2,...,K\}$. Then,

$x_i(\eta_i(\omega_i), m_{-i}) = -\hat{\varepsilon}$ for all $m_{-i} \in M_{-i}$.

Since $x_i(m) \geq -\hat{\varepsilon}$, $g(m) = g(\eta_i(\omega_i), m_{-i})$ for all $m_{-i} \in M_{-i}$, and agent $i$ has preference for honesty, it follows that agent $i$ has incentive to announce $m_i$ instead of $\eta_i(\omega_i)$. Next, suppose that $\eta_i^k(\omega_i) = \eta_i^{\hat{K}+1}(\omega_i)$ for all $k \in \{\hat{K}+2,...,K\}$. By replacing $m_i$ with $\eta_i(\omega_i)$, we can decrease the realized cost for dishonest reporting by $(\frac{K-\hat{K}}{K})c$. Since $x_i(m) - x_i(\eta_i(\omega_i), m_{-i}) \geq -\hat{\varepsilon}$ and $g(m) = g(\eta_i(\omega_i), m_{-i})$ for all $m_{-i} \in M_{-i}$, it follows from the inequality (2) that agent $i$ has incentive to announce $m_i$ instead of $\eta_i(\omega_i)$. Here, the inequality (2) implies that the cost for making the last $K-\hat{K}$ announcements dishonestly, i.e., $\frac{K-\hat{K}}{K}c$, is larger than the monetary fine $\varepsilon$. This is the driving force of incentivizing each agent to reporting honestly from the $(\hat{K}+1)-th$ component to the $K-th$ announcement.

Fix $\eta \in \Xi$ and $k \in \{1,...,\hat{K}\}$ arbitrarily, where $\eta_i^h(\omega_i) = \hat{\eta}_i^h(\omega_i)$ for all $i \in N$, all $\omega_i \in \Omega_i$, and all $h \in \{1,...,k-1\} \bigcup \{\hat{K}+1,...,K\}$. Fix $i \in N$, and $\omega_i \in \Omega_i$ arbitrarily, and suppose $\eta_i^k(\omega_i) \neq \hat{\eta}_i^k(\omega_i)$. Let $m_i \in M_i$ denote the message for agent $i$ defined by

$$m_i^k = \hat{\eta}_i^k(\omega_i),$$

and

$$m_i^h = \eta_i^h(\omega_i) \text{ for all } h \in \{1,...,K\}/\{k\}.$$

Let $\widetilde{\Omega}_{-i} \subset \Omega_{-i}$ denote the set of private signal profiles $\omega_{-i}$ for the other agents satisfying that $\eta_j^k(\omega_j) \neq \hat{\eta}_j^k(\omega_j)$ for some $j \in N/\{i\}$. Note that for every $\omega_{-i} \in \widetilde{\Omega}_{-i}$,

$$x_i(\eta(\omega)) = -\hat{\varepsilon} \text{ and } x_i(m_i, \eta_{-i}(\omega_{-i})) = 0.$$

Hence,

$$\begin{aligned}
&E[u_i(g(\eta(\omega)), x_i(\eta(\omega)), \eta_i(\omega_i), \omega) \mid p, \omega_i] \\
&- E[u_i(g(m_i, \eta_{-i}(\omega_{-i})), x_i(m_i, \eta_{-i}(\omega_{-i})), m_i, \omega) \mid p, \omega_i] \\
&\leq \frac{1}{\hat{K}}\{E[v_i(f(\eta^k(\omega)), \omega) \mid p, \omega_i] - E[v_i(f(\omega_i, \eta_{-i}^k(\omega_{-i})), \omega) \mid p, \omega_i]\} - \frac{c}{K} \\
&- \hat{\varepsilon} \sum_{\omega_{-i} \in \widetilde{\Omega}_{-i}} p_i(\omega_{-i} \mid \omega_i) \\
&= \frac{1}{\hat{K}}\{E[v_i(f(\eta_i^k(\omega_i), \omega_{-i}), \omega) \mid p, \omega_i] - E[v_i(f(\omega), \omega) \mid p, \omega_i]\} - \frac{c}{K} \\
&+ \sum_{\omega_{-i} \in \widetilde{\Omega}_{-i}} [\frac{1}{\hat{K}}\{v_i(f(\eta^k(\omega)), \omega) - v_i(f(\eta_i^k(\omega_i), \omega_{-i}), \omega) \\
&+ v_i(f(\omega), \omega) - v_i(f(\omega_i, \eta_{-i}^k(\omega_{-i})), \omega)\} - \hat{\varepsilon}] p_i(\omega_{-i} \mid \omega_i) \\
&\leq \frac{1}{\hat{K}}\{E[v_i(f(\eta_i^k(\omega_i), \omega_{-i}), \omega) \mid p, \omega_i] - E[v_i(f(\omega), \omega) \mid p, \omega_i]\} - \frac{c}{K}
\end{aligned}$$

$$+\{\frac{2}{\hat{K}}\max_{(a,a',\omega,i)\in A^2\times\Omega\times N}\left|v_i(a,\omega)-v_i(a',\omega)\right|-\hat{\varepsilon}\}\sum_{\omega_{-i}\in\tilde{\Omega}_{-i}}p_i(\omega_{-i}\mid\omega_i),$$

which is less than zero because of the inequalities (3) and (4).

The above arguments imply that $\hat{\eta}$ is the unique Bayesian iteratively undominated message rule profile. Hence, we have proved that $f$ is implemented by $\hat{G}(f,K,\hat{K},\hat{\varepsilon},c)$ in Bayesian Nash equilibrium for $p$.

**Q.E.D.**

We must note that for every $c>0$ and every $\varepsilon>0$, there exist $\hat{\varepsilon}\in(0,\varepsilon]$, $K$, and $\hat{K}\in\{1,...,K-1\}$ such that the inequalities (2) and (3) hold and $\frac{\hat{K}}{K}$ is close to unity. Hence, it follows from Theorem 2 that *for every $c>0$ and every $\varepsilon>0$, there exist $\hat{\varepsilon}\in(0,\varepsilon]$, $K$, and $\hat{K}\in\{1,...,K-1\}$ such that for every $p\in P$, every social choice function $f\in F$ is implemented in Bayesian Nash equilibrium by $\hat{G}(f,K,\hat{K},\hat{\varepsilon},c)$ if for every $i\in N$ and every $\omega_i\in\Omega_i$,*

*(5)* $E[v_i(f(\omega),\omega)\mid p,\omega_i]>E[v_i(f(\omega_i',\omega_{-i}),\omega)\mid p,\omega_i]-c$ *for all* $\omega_i'\in\Omega_i$.

Here, the inequalities (5) imply that the social choice function $f$ is *Bayesian incentive compatible associated with the cost c for dishonest reporting*. This positive result is in contrast with the standard implementation models with incomplete information, where each agent has preference only for consequences, i.e., $c=0$, and Bayesian incentive compatibility is necessary but not sufficient for implementation in Bayesian Nash equilibrium. This negative result is valid, even if we replace Bayesian Nash equilibrium with any other equilibrium concept with incomplete information and consider virtual implementation.[12]

The proof of Theorem 2 shows that $\hat{\eta}$ is the unique Bayesian iteratively undominated message rule profile in $\hat{G}(f,K,\hat{K},\hat{\varepsilon},c)$ for $p$. This implies that $f$ is implementable not only in mixed Bayesian Nash equilibrium but also in Bayesian iteratively undominated strategies.

We must note that the mechanism $\hat{G}(f,K,\hat{K},\hat{\varepsilon},c)$ is *universal* in the sense that it does not depend on the detail of the private signal structure $p\in P$. This is in contrast with the standard implementation models with incomplete information, where the constructed mechanisms depend crucially on the fine details of $p$, and therefore, might be difficult to use in practice.

A social choice function $f\in F$ is said to be *universally* implemented by a mechanism $G$ in Bayesian Nash equilibrium if it is implemented by $G$ in Bayesian Nash equilibrium for all private signal structures.

---

[12] See Jackson (1991), Matsushima (1993), Duggan (1997), Serrano and Vohra (2000), and others.

**Theorem 3:** *Suppose that the inequalities (2) and (3) hold. Then, every social choice function* $f \in F$ *is universally implemented by* $\hat{G}(f, K, \hat{K}, \hat{\varepsilon}, c)$ *in Bayesian Nash equilibrium if for every* $i \in N$ *and every* $\omega \in \Omega$,

(6) $$v_i(f(\omega), \omega) > v_i(f(\omega_i', \omega_{-i}), \omega) - \frac{\hat{K}}{K} c \text{ for all } \omega_i' \in \Omega_i .$$

**Proof:** The inequalities (6) imply the inequalities (4) for all $p \in P$. Hence, Theorem 2 implies Theorem 3.

**Q.E.D.**

We must note from Theorem 3 and the above arguments that for every $c > 0$ and every $\varepsilon > 0$, there exist $\hat{\varepsilon} \in (0, \varepsilon]$, $K$, and $\hat{K} \in \{1, ..., K-1\}$ such that every social choice function $f \in F$ is universally implemented by $\hat{G}(f, K, \hat{K}, \hat{\varepsilon}, c)$ in Bayesian Nash equilibrium if for every $i \in N$ and every $\omega \in \Omega$,

(7) $$v_i(f(\omega), \omega) > v_i(f(\omega_i', \omega_{-i}), \omega) - c \text{ for all } \omega_i' \in \Omega_i .$$

Here, the inequalities (7) imply that $f$ is *ex post incentive compatible with the cost c for dishonest reporting,* i.e., the honest message rule profile $\hat{\eta}$ is an *ex post equilibrium* originated by Cremer and McLean (1985) in the direct mechanism with no fines. Hence, it follows that *for every* $c > 0$, *every* $\varepsilon > 0$, *and every* $f \in F$, *ex post incentive compatibility implied by the inequalities (7) is a sufficient condition for implementation in Bayesian Nash equilibrium via a universal mechanism in the above sense*.

## References


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

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

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

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

Alger, I. and C. A. Ma (2003): "Moral Hazard, Insurance, and Some Collusion," *Journal of Economic Behavior and Organization* 50, 225-247.

Basu, K., P. Pattanaik, and K. Suzumura (1995): *Choice, Welfare, and Development*, Oxford: Clarendon Press.

Cremer, J. and R. McLean (1985): "Optimal Selling Strategies under Uncertainty for a Discriminating Monopolist When Demands are Interdependent," *Econometrica* 53, 345-361.

Deneckere, R. and S. Severinov (2001): "Mechanism Design and Communication Costs," mimeo.

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

Dworkin, R. (1981): "What is equality ? Part 1: Equality of Welfare, Part 2: Equality of Resources," *Philosophy and Public Affairs* 10, 185-246, 283-345.

Eliaz, K. (2001): "Fault Tolerant Implementation," forthcoming in *Review of Economic Studies*.

Erard, B. and J. Feinstein (1994): "Honesty and Evasion in the Tax Compliance," *RAND Journal of Economics* 25, 1-19.

Glazer, J. and A. Rubinstein (1998): "Motives and Implementation: On the Design of Mechanisms to Elicit Options," *Journal of Economic Theory* 79, 157-173.

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

Maskin, E. and T. Sjostrom (2002): "Implementation Theory," in *Handbook of Social Choice and Welfare, Volume 1*, ed. by K. Arrow, A. Sen, and K. Suzumura, Elsevier Science.

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

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

Moore, J. (1992): "Implementation, Contracts, and Renegotiation in Environments with Complete Information," in *Advances in Economic Theory: Sixth World Congr*ess, ed. by J.-J. Laffont, Cambridge University Press.

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

Rawls, J. (1971): *A Theory of Justice*, Cambridge: Harvard University Press.
Sen, A. (1982): *Choice, Welfare and Measurement*, Oxford: Blackwell.
Sen, A. (1985): *Commodities and Capabilities*, Amsterdam: North-Holland.
Sen, A. (1999): "The Possibility of Social Choice," *American Economic Review* 89, 349-378.
Serrano, R. and R. Vohra (2000): "Type Diversity and Virtual Bayesian Implementation," Working Paper No. 00-16, Department of Economics, Brown University.
Serrano, R. and R. Vohra (2001): "Some Limitations of Virtual Bayesian Implementation," *Econometrica* 69, 785-792.
Suzumura, K. (2002): "Introduction," in *Handbook of Social choice and Welfare, Volume 1*, ed. by K. Arrow, A. Sen, and K. Suzumura, Elsevier Science.