CIRJE-F-178

Honesty-Proof Implementation

Hitoshi Matsushima
University of Tokyo

October 2002

Honesty-Proof Implementation

Hitoshi Matsushima[+]

Faculty of Economics, University of Tokyo

March 4, 2002
This Version: October 10, 2002

Abstract

We investigate implementation of social choice functions that map from states to lotteries and may depend on factors other than agents' preferences. We assume that agents are not only purely self-interested but also honesty-oriented in a lexicographical way. We define iterative honesty-proofness by iteratively removing messages dominated by more honest messages. We show that in the complete information environments with small fines, every social choice function is implementable in iterative honesty-proofness. This is in contrast with the standard implementation model, because any 'normative' social choice function depending on non-preference factors is never implementable when agents are not influenced by factors other than pure self-interest. We extend this result to the incomplete information environments with quasi-linearity and with correlated private signals.

Next, we assume that it is costly for each agent to report dishonestly and this cost may be close to zero. We show that in the incomplete information environments, every incentive compatible social choice function can be implemented by the mechanism that is universal in the sense that it does not depend on the private signal structure.

Keywords: Honesty-Orientation, Normative Social Choice Functions, Implementation, Small Fines, Universal Mechanisms.

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

1. Introduction

We investigate implementation of social choice functions that map from states to lotteries over the set of pure alternatives and may depend on factors other than agents' preferences. We will take into account the fact that real human behavior may be influenced by factors other than pure self-interest such as *honesty-orientation*.[1] The purpose of this paper is to show that, by allowing agents to be honesty-oriented only in marginal ways, we can drastically expand the class of implementable social choice functions. We also discuss the possibility of implementing a social choice function by using a mechanism that does not depend on the detail of the model structure.

First, we investigate the complete information environments where there exist three agents who know what is the true state. We assume that there exists a single agent who is not only purely self-interested but also honesty-oriented in a *lexicographical* way. That is, this agent prefers a message to another message if the former is more honest than the latter and both messages provide her with the same expected utility. We define the solution concept named *iterative honesty-proofness* by iteratively removing messages that are dominated by its more honest messages. We require a mechanism to have the *unique* iteratively honesty-proof message profile at every state. We allow the central planner to fine agents only *small* amounts of the private goods.

We show that every social choice function is implementable in iterative honesty-proofness. This is in sharp contrast with the standard model of implementation where all agents are never influenced by factors other than pure self-interest.[2] A social choice function is said to be *normative* if there exist distinct states that provide agents with the same preference profile, but to that it assigns different lotteries. We must note that, in the standard model, it is impossible for any normative social choice function to be implementable.[3] For instance, we require a mechanism to have the unique Nash equilibrium outcome at every state, and require this outcome to be equivalent to the lottery assigned by the social choice function to this state. Whenever a pair of distinct states induces the same preference profile, then the agents are faced with the same game between these states. Hence, the resultant sets of Nash equilibria must be equivalent between these states. This, however, is a contradiction because there exists such a pair of states to that the normative social choice function assigns different lotteries.[4]

[1] There exist works on psychological games such as Geanakoplos, Pearce, and Stacchetti (1989) and Rabin (1993) in which players are not purely self-interested. For surveys on psychology and economics, see Rabin (1998) and Fehr and Schmidt (2001).

[2] There exist recent works in the implementation literature such as Eliaz (2001) that take into account factors other than pure self-interest such as bounded rationality. These works checked the robustness of implementability in terms of ideally rational equilibrium concepts.

[3] For the survey of implementation in the complete information environments, see Moore (1992). Most authors defined a state as the preference profile of agents, which automatically excludes the class of all normative social choice functions.

[4] There exist many works in implementation with complete information that showed their respective possibility theorems by weakening the requirement of Nash implementability. For instance, Moore and

A state as the input of a social choice function may involve information about factors that are not relevant to agents' preferences. Whenever a social choice function depends on such factors, then it is normative. Social choice functions that make use of interpersonal utility comparison such as the utilitarian social choice function (D'Aspremont and Gevers (1977)) and the leximin social choice function (Deschamps and Gevers (1978)) are all normative. Economic judgments in real situations are sometimes based on non-utility factors such as poverty, inequality, and rights. Several philosophers and ethical economists emphasized that such non-utility factors do play the crucial role in ethical judgments and introduced their respective notions of anti-welfarism such as primary goods (Rawls (1971)) and capabilities (Sen (1985)).[5]

Even from more positivist-like viewpoints, normative social choice functions are very important to investigate. Suppose that there exist disadvantageous individuals who cannot participate in the decision procedure but are deeply influenced by the decision. Then, the impact of these individuals' preferences on the ethical judgment should be crucial. Hence, the social choice function must depend on these individuals' preferences. The participants' preferences, however, do not well represent their preferences. In this case, the social choice function must be normative in our sense. Moreover, Serrano and Vohra (1998) presented an example (Example I in their paper) in the economic environments with incomplete information, in which all agents' preferences are the same across states but their initial endowments depend on the state. In their example, any individually rational social choice function is inconstant. Hence, every individually rational social choice function must be normative.

From the observations above, it is clear that a social choice function may be normative when the set of states as its domain is inclusive and it reflects the equity (or individual rationality) in a society. In spite of its apparent importance, however, it is a widely held view in implementation theory that no normative social choice function can be implemented through decentralized decision making in the complete information environments. In contrast, the result of this paper implies that we can make every normative social choice function implementable merely by introducing the honesty-oriented motive of a single agent that is weak enough to be *consistent* with this agent's purely self-interested motive.

Next, we investigate the incomplete information environments where agents have their private signals regarding what is the true state. We assume that there exist three or more agents, and all agents are not only purely self-interested but also honesty-oriented in the

Repullo (1988) and Palfrey and Srivastava (1991) replaced Nash equilibrium with weaker solution concepts such as subgame perfect equilibrium and undominated Nash equilibrium. These works did not take the class of normative social choice functions into account. In fact, no normative social choice function is implementable even if we replace Nash equilibrium with any such purely self-interested solution concept.

[5] Classical social choice theory assumed that a social welfare functional depends only on agents' preference profile, and therefore, did not investigate normative social choice in our sense. Sen criticized classical social choice theory for allowing only such extremely narrow informational bases. See Sen (1982, 1999).

lexicographical way. We define the solution concept named *Bayesian iterative honesty-proofness*, require a mechanism to have the unique Bayesian iteratively honesty-proof message profile at every state, and allow the central planner to fine agents only small amounts. We assume that utilities are quasi-linear. We also assume that agents' private signals are *correlated*.

We show that every incentive compatible social choice function is implementable in Bayesian iterative honesty-proofness. This is in contrast with Bayesian Nash implementation as follows. Suppose that all agents have complete knowledge about their preference profile. Then, it is impossible for any normative social choice function to be implementable in Bayesian Nash equilibrium. Serrano and Vohra (2001), for instance, showed in their example that with complete knowledge, any individually rational social choice function is not implementable in Bayesian Nash equilibrium, even if it is incentive compatible. In contrast, every normative and incentive compatible social choice function is implementable in Bayesian iterative honesty-proofness, even if agents have complete knowledge about their preferences.[6]

Finally, we reconsider the incomplete information environments where there may exist only two agents. We assume that it is *costly* for each agent to report dishonestly for psychological reasons. Hence, all agents are not purely self-interested. Based on this, we introduce a modified version of Bayesian iterative honesty-proofness.

Several past works such as Erard and Feinstein (1994), Alger and Ma (1998), and Deneckere and Severinov (2001) examined the case that agents' ability to manipulate information is limited, and demonstrated that including honesty-oriented agents could significantly alter the model. These works commonly assumed that the cost of reporting dishonestly is so large that we may not need to impose the standard incentive compatibility constraints on an implementable allocation.[7] In contrast, this paper allows the maximal total cost of reporting dishonestly to be as close to zero as possible, and focuses on the uniqueness constraints instead of the incentive compatibility constraints. Hence, we can say that every agent is, not exactly, but virtually, purely self-interested.

We show that every incentive compatible social choice function is implementable. This result is very permissive as follows. We do not require any restriction on the private signal structure such as whether agents' private signals to be correlated or independent. Of particular importance, the constructed mechanism is *universal* in the sense that it does not depend on the private signal structure. Hence, the central planner can construct the mechanism without any knowledge about the private signal structure. Moreover, whenever

[6] For the survey of implementation in the incomplete information environments, see Palfrey (1992). When agents do not have complete knowledge about their preferences, there may exist normative social choice functions that are implementable in Bayesian Nash equilibrium. For characterization and permissive results, see Jackson (1992), Matsushima (1993), Abreu and Matsushima (1992b), Duggan (1997), and Serrano and Vohra (2000).

[7] Deneckere and Severinov (2001) assumed that the cost of reporting a single dishonest message is small. They constructed mechanisms in which each agent can make multiple announcements. Here, the total cost of reporting multiple dishonest messages may be large enough to make the standard incentive compatibility constraints needless.

truth telling is an ex post equilibrium in the revelation game, then even agents do not need to know the private signal structure. These points are in sharp contrast with the previous works in Bayesian implementation where a mechanism was tailored to a specific situation, and therefore, it would be difficult to use in practice.[8]

In most part of this paper, we construct mechanisms where each agent makes multiple announcements. The paper is related to Abreu and Matsushima (1992a, 1992b, 1994) in this respect. In contrast to Abreu and Matsushima, however, the paper does not use the device of virtualness originated in Matsushima (1988, 1993) and Abreu and Sen (1991). Hence, our possibility theorems are on 'exact' implementation as opposed to 'virtual', and our mechanisms might be much simpler than the mechanisms in Abreu and Matsushima for this reason.[9]

The organization of the paper is as follows. Section 2 defines the basic model and introduces normative social choice functions. Section 3 introduces honesty-orientation in the lexicographical sense, and shows that in the complete information environments, every social choice function is implementable in iterative honesty-proofness with small fines. Section 4 constructs so-called *modulo mechanisms*, and shows that even without fines, many normative social choice functions are implementable when we replace iterative honesty-proofness with a weaker concept named *honesty-proof Nash equilibrium*. We also argue on a complementary role of *salience-orientation*. Section 5 shows that in the incomplete information environments, every incentive compatible social choice function is implementable in Bayesian honesty-proofness when agents' private signals are correlated. Section 6 assumes that it is costly for every agent to report dishonestly, and shows that even if this cost is close to zero, every incentive compatible social choice function can be implemented by the universal mechanism. Finally, Section 7 concludes.

[8] A relevant point can be found in auction theory where from a practical standpoint a mechanism is sometimes restricted not to depend on the fine detail of the private signal structure. In the auction theory context, the terminology of 'universal' may have more restrictive implications than in this paper. For instance, we may require an auction-like mechanism not to depend on the social choice function. See Krishna (2002).

[9] Abreu and Matsushima (1994) studied exact implementation but used a similar idea of virtualness to Abreu and Matsushima (1992a).

2. The Model

Let $N=\{1,...,n\}$ denote the finite set of agents whom the planner requires to announce messages.[10] Let Ω and A denote the nonempty and compact sets of states and pure alternatives, respectively. The utility function for each agent $i\in N$ is given by

$$u_i : A\times R\times\Omega\to R.$$

The central planner will choose a pure alternative and fine each agent a nonnegative amount of the private goods that is less than or equals

$$\varepsilon+\xi\geq 0,$$

where ε and ξ are nonnegative real numbers. When the central planner chooses $a\in A$ and fines each agent $i\in N$ the amount $-t_i\in[0,\varepsilon+\xi]$ at the state $\omega\in\Omega$, the resultant utility for agent i is given by $u_i(a,t_i,\omega)$. We assume that $u_i(a,t_i,\omega)$ is increasing with respect to t_i for every $(a,t_i,\omega)\in A\times[-\varepsilon-\xi,0]\times\Omega$.

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 Δ denote the set of all simple lotteries. We assume the expected utility hypothesis, where the expected utility for agent i when the central planner chooses a pure alternative according to $\alpha\in\Delta$ and fines agent i the amount $-t_i$ at the state $\omega\in\Omega$ is denoted by $u_i(\alpha,t_i,\omega)\equiv\sum_{a\in A}u_i(a,t_i,\omega)\alpha(a)$.

A *social choice function* is defined as a mapping from states to simple lotteries and is denoted by $f:\Omega\to\Delta$, where $f(\omega)\in\Delta$ is regarded as the socially desired simple lottery at the state $\omega\in\Omega$. We denote by F the set of all social choice functions.

A pair of distinct states $\omega\in\Omega$ and $\omega'\in\Omega/\{\omega\}$ is said to be *preference-equivalent* if these states induce the same preference profile, i.e., if for every $i\in N$, there exist $\beta_i>0$ and $\gamma_i\in R$ such that

$$u_i(a,t_i,\omega')=\beta_i u_i(a,t_i,\omega)+\gamma_i \text{ for all } a\in A \text{ and all } t_i\in[-\varepsilon-\xi,0].$$

A social choice function $f\in F$ is said to be *normative* if there exists a preference-equivalent pair of states $\omega\in\Omega$ and $\omega'\in\Omega/\{\omega\}$ such that

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

Here, different simple lotteries may be socially desired even if the agents have the same preference profile.

A *mechanism* is defined by $G=(M,g,x)$, where M_i is the set of messages for each agent $i\in N$, $M=\prod_{i\in N}M_i$, $g:M\to\Delta$, $x=(x_i)_{i\in N}$, and $x_i:M\to[-\varepsilon-\xi,0]$. When the agents announce the message profile $m\in M$, the central planner chooses a pure alternative

[10] There may exist other individuals who are relevant to the central planner's decision but do not participate in the decision procedure.

according to the simple lottery $g(m) \in \Delta$ and fines each agent $i \in N$ the amount $-x_i(m) \in [0, \varepsilon + \xi]$.

A message profile $m \in M$ is said to be a *Nash equilibrium* in the game defined by (G, ω) if for every $i \in N$, and every $m_i' \in M_i$,

$$u_i(g(m), x_i(m), \omega) \geq u_i(g(m/m_i'), x_i(m/m_i'), \omega).$$

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

$$g(m) = f(\omega).$$

Proposition 1: *For every normative social choice function* f*, there exists no mechanism* G *that implements* f *in Nash equilibrium.*

Proof: Let a pair of states $\omega \in \Omega$ and $\omega' \in \Omega / \{\omega\}$ be preference-equivalent and satisfy that $f(\omega) \neq f(\omega')$. Whenever a message profile m is a Nash equilibrium in (G, ω), then it is also a Nash equilibrium in (G, ω'). Hence, both $g(m) = f(\omega)$ and $g(m) = f(\omega')$ must hold if f is implemented by G in Nash equilibrium. This, however, is a contradiction because $f(\omega) \neq f(\omega')$.

Q.E.D.

In the same way as in Proposition 1, it follows that no normative social choice function is implementable in any purely self-interested solution concept in the complete information environments such as dominant strategies, iterative dominance, undominated Nash equilibrium, perfect Nash equilibrium, and strict Nash equilibrium. Moreover, it follows that no normative social choice function is virtually implementable in any purely self-interested solution concept in the complete information environments, because any other social choice function that is sufficiently close to the normative social choice function must be normative. In Section 5, we will show that even in the incomplete information environments, a similar impossibility result may hold.

3. Complete Information with Small Fines

This section considers the *complete* information environments where all agents know what is the true state. We assume that only *three* agents are required to announce messages, i.e.,

$$n = 3.$$[11]

We allow the central planner to fine agents positive amounts of the private goods, i.e.,

$$\varepsilon > 0 \text{ and } \xi > 0,$$

both of which may be close to zero.

We consider only a class of mechanisms $G^* = (M, g, x)$ satisfying that for every $i \in N$,

$$M_i = \Omega^{K_i},$$

where K_i is a positive integer. Each agent $i \in N$ announces K_i elements of Ω at one time as her multiple opinions about the state. For every $i \in N$, let $M_i = M_i^1 \times \cdots \times M_i^{K_i}$ and $m_i = (m_i^1, ..., m_i^{K_i})$, where $M_i^k = \Omega$ and $m_i^k \in M_i^k$. The *honest message for each agent* $i \in N$ *at each state* $\omega \in \Omega$ is defined as $\mu_i(\omega) = (\mu_i^k(\omega))_{k=1}^{K_i} \in M$ where

$$\mu_i^k(\omega) = \omega \text{ for all } k \in \{1, ..., K_i\}.$$

Let $\mu(\omega) = (\mu_i(\omega))_{i \in N}$ denote an honest message profile. We further confine our attention to mechanisms satisfying that there exists a positive integer K such that

$$K_1 = K + 1 \text{ and } K_2 = K_3 = K.$$

Agent 1 announces $K+1$ elements of Ω, whereas each of agents 2 and 3 announces K elements of Ω. For every $k \in \{1, ..., K\}$, let $m^k = (m_1^k, ..., m_n^k)$ denote the profile of the $k-th$ opinions.

3.1. Iterative Honesty-Proofness

This section assumes that only agent 1 is *honesty-oriented* in a lexicographical sense that she prefers m_1 to m_1' if m_1 is the same as m_1' except for the $(K+1)-th$ opinion, both m_1 and m_1' provide her with the same expected utility, and m_1 induces agent 1 to make the honest announcement as her $(K+1)-th$ opinion, i.e., $m_1^{K+1} = \mu_1^{K+1}(\omega) \neq m_1'^{K+1}$. Hence, agent 1 is required to be honesty-oriented *only* for her $(K+1)-th$ announcement.

We introduce the solution concept named *iterative honesty-proofness* as follows. Let $M_1(1, \omega) \subset M_1$ denote the set of all messages m_1 for agent 1 satisfying that either $m_1^{K+1} = \mu_1^{K+1}(\omega)$, or

$$u_1(g(m), x_1(m), \omega) > u_1(g(m/m_1'), x_1(m/m_1'), \omega) \text{ for some } m_{-1} \in M_{-1},$$

where $m_1' = (m_1^1, ..., m_1^K, \mu_1^{K+1}(\omega))$ is the message for agent 1 defined by replacing m_1^{K+1} with

[11] This implies that all other individuals have single-valued sets of messages. We do not require all other individuals to know the true state.

$\mu_1^{K+1}(\omega)$. This first round of iterative removal implies that agent 1 is honesty-oriented in the sense that for every $m \in M$, whenever $m_1^{K+1} \neq \mu_1^{K+1}(\omega)$ and m_1' always provides her with at least the same expected utility as m_1 irrespective of the other agents' messages, then she never announces m_1. For every $i \in N/\{1\}$, let $M_i(1,\omega) \equiv M_i$. For every integer $h \geq 1$, let $M(h,\omega) \equiv \prod_{i \in N} M_i(h,\omega)$, and let $M_{-i}(h,\omega) \equiv \prod_{j \in N/\{i\}} M_j(h,\omega)$ for all $i \in N$. For every $h \geq 2$, and every $i \in N$, let $M_i(h,\omega)$ denote the set of all messages $m_i \in M_i$ for agent i satisfying that $m_i \in M_i(h-1,\omega)$, and there exists no $m_i' \in M_i(h-1,\omega)$ such that

$$u_i(g(m/m_i'), x_i(m/m_i'), \omega) > u_i(g(m), x_i(m), \omega) \text{ for all } m_{-i} \in M_{-i}(h-1,\omega).$$

Hence, in every round of iterative removal except the first round, we require only that each agent is purely self-interested in the sense that she never announces strictly dominated messages. Let

$$M_i(\infty,\omega) \equiv \lim_{h \to \infty} M_i(h,\omega) \text{ and } M(\infty,\omega) \equiv \prod_{i \in N} M_i(\infty,\omega).$$

A message profile $m \in M$ is said to be *iteratively honesty-proof* in the game (G^*,ω) if

$$m \in M(\infty,\omega).$$

Note that the definition of iterative honesty-proofness is irrelevant to the order of iterative removal, provided that agent 1 is honesty-oriented not only in the first round but also in the other rounds. Note also that if there exists the unique iteratively honesty-proof message profile $m \in M$ in (G^*,ω), then it is the unique Nash equilibrium in (G^*,ω) satisfying that agent 1 behaves as being honesty-oriented in the sense that either $m_1^{K+1} = \mu_1^{K+1}(\omega)$, or

$$u_1(g(m), x_1(m), \omega) > u_1(g(m/m_1'), x_1(m/m_1'), \omega),$$

where $m_1' = (m_1^1,\ldots,m_1^K,\mu_1^{K+1}(\omega))$. A social choice function f is said to be *implemented by the mechanism* G^* *in iterative honesty-proofness* if for every $\omega \in \Omega$, the honest message profile $\mu(\omega) \in M$ is the unique iteratively honesty-proof message profile in (G^*,ω),

$$g(\mu(\omega)) = f(\omega),$$

and

$$x_1(\mu(\omega)) = x_2(\mu(\omega)) = x_3(\mu(\omega)) = 0.$$

3.2. Specification of Mechanisms

We specify a mechanism $G^* = G^*(f,K)$ as follows. Fix a lottery $\overline{\alpha} \in \Delta$ arbitrarily. We define $z : \Omega^3 \to \Delta$ in ways that for every $\omega \in \Omega$, and every $\omega' \in \Omega$,

$$z(\omega,\omega,\omega') = z(\omega,\omega',\omega) = z(\omega',\omega,\omega) = f(\omega),$$

and for every $\omega'' \in \Omega$,

$$z(\omega,\omega',\omega'') = \overline{\alpha} \text{ if } \omega \neq \omega' \neq \omega'' \neq \omega.$$

We may regard z as the *majority rule* in the sense that to every triplicate of opinions, it assigns the simple lottery that is socially desired when its majority opinion is correct. We specify g by

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

For every $k\in\{1,...,K\}$, with probability $\frac{1}{K}$, the central planner chooses a pure alternative according to $z(m^k)\in\Delta$.

For every $i\in N$, and every $m\in M$, we denote by $k(i,m)\in\{0,...,K\}$ the number of integers $k\in\{1,...,K\}$ satisfying that there exists $\omega\in\Omega$ such that ω is the k-th majority opinion that is announced by the other agents than agent 1, i.e., $m_j^k=\omega$ for all $j\in N/\{i\}$, and agent i does not announce this opinion, i.e., $m_i^k\neq\omega$. Hence, each agent $i\in N$ announces $k(i,m)$ opinions that are different from their respective majority opinions when $m\in M$ is announced. We specify x_1 in ways that for every $m\in M$,

$$x_1(m)=-\frac{k(1,m)\xi}{K}.$$

For every $i\in N/\{1\}$, we specify x_i in ways that for every $m\in M$,

$$x_i(m)=-\varepsilon-\frac{k(i,m)\xi}{K} \text{ if there exists } k\in\{1,...,K\} \text{ such that } m_i^k\neq m_1^{K+1} \text{ and}$$
$$m_2^h=m_3^h=m_1^{K+1} \text{ for all } h\in\{1,...,k-1\},$$

and

$$x_i(m)=-\frac{k(i,m)\xi}{K} \text{ otherwise.}$$

Each of agent 2 and agent 3 is fined the amount ε if and only if she is the first to announce a different opinion from agent 1's (K+1)-th opinion. For every $k\in\{1,...,K\}$, each agent is fined the amount $\frac{\xi}{K}$ if and only if there exists the k-th majority opinion and she does not announce it.

Our construction is similar to Abreu and Matsushima (1992a, 1994) in that each agent announces multiple opinions about the state, and the central planner fines the first deviants. Our construction, however, is much simpler than Abreu and Matsushima, because we do not use the device of virtualness originated in Matsushima (1988) and Abreu and Sen (1991).[12]

When all agents announce the honest message profile $\mu(\omega)$, the central planner chooses a pure alternative according to the lottery $f(\omega)$ that is assigned by the social choice function f to the true state $\omega\in\Omega$ and no agents are fined, i.e., for every $\omega\in\Omega$,

$$g(\mu(\omega))=f(\omega) \text{ and } x_1(\mu(\omega))=x_2(\mu(\omega))=x_3(\mu(\omega))=0.$$

[12] Abreu and Matsushima (1992a) did not assume that each agents can be fined a small amount of the private goods, and instead did assume a weaker restriction that the central planner can fine each agent by changing an allocation to another one satisfying that she prefers the former to the latter. Abreu and Matsushima (1994) assumed that utilities are quasi-linear, whereas the present paper does not assume the quasi-lineality except Sections 5 and 6.

3.3. The Possibility Theorem

Note that for every $\omega \in \Omega$, and every $\omega' \in \Omega / \{\omega\}$, the message profile $\mu(\omega')$ is a Nash equilibrium in the game $(G^*(f,K),\omega)$, because for every $i \in N$, and every $m_i \in M_i$,

$$u_i(g(\mu(\omega')/m_i), x_i(\mu(\omega')/m_i), \omega) = u_i(f(\omega'), x_i(\mu(\omega')/m_i), \omega)$$
$$\leq u_i(f(\omega'), 0, \omega) = u_i(g(\mu(\omega')), x_i(\mu(\omega')), \omega).$$

This implies that every inconstant social choice function $f \in F$ is never implemented by the mechanism $G^*(f,K)$ in Nash equilibrium, irrespective of K. In contrast, the following theorem states that whenever K is chosen sufficiently large then every social choice function $f \in F$ can be implemented by $G^*(f,K)$ in iterative honesty-proofness.

Theorem 2: *For every social choice function $f \in F$, there exists K such that f is implemented by $G^*(f,K)$ in iterative honesty-proofness.*

Proof: Since A and Ω are compact, there exist positive real numbers $\rho > 0$ and $v > 0$ such that for every $i \in N$, and every $(a, t_i, \omega) \in A \times [-\varepsilon - \xi, 0] \times \Omega$,

(1) $$\lim_{t_i' \uparrow t_i} \frac{u_i(a,t_i,\omega) - u_i(a,t_i',\omega)}{t_i - t_i'} \geq \rho, \text{ and}$$

(2) $$|u_i(a,t_i,\omega) - u_i(a',t_i,\omega)| \leq v \text{ for all } a' \in A.$$

Choose K so as to satisfy

(3) $$K > \frac{v}{\rho\varepsilon}.$$

Fix $\omega \in \Omega$ arbitrarily. Agent 1 has incentive to announce $m_1^{K+1} = \mu_1^{K+1}(\omega)$ in the game $(G^*(f,K),\omega)$, because both $g(m)$ and $x_1(m)$ are independent of m_1^{K+1}. Fix $k \in \{1,...,K\}$ and $m \in M$ arbitrarily, and suppose that

$$m_1^{K+1} = \mu_1^{K+1}(\omega),$$

and

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

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

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

and

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

Suppose $i \neq 1$. If $m_j^k = \mu_j^k(\omega)$ for all $j \in N/\{i\}$, then, $g(m)$ is independent of m_i^k, and

$$x_i(m/m_i') - x_i(m) \geq \frac{\xi}{K} > 0,$$

which 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

$$x_i(m/m_i') - x_i(m) = \varepsilon + \frac{\xi}{K},$$

and therefore, agent i has incentive to announce m_i' instead of m_i, because

$$\begin{aligned}
&u_i(g(m),t_i(m),\omega) - u_i(g(m/m_i'),t_i(m/m_i'),\omega)\\
&= u_i(g(m),t_i(m),\omega) - u_i(g(m/m_i'),t_i(m),\omega)\\
&+ \{u_i(g(m/m_i'),t_i(m),\omega) - u_i(g(m/m_i'),t_i(m/m_i'),\omega)\}\\
&= \frac{1}{K}\{u_i(z(m^k),t_i(m),\omega) - u_i(z(m^k/m_i'^k),t_i(m),\omega)\}\\
&+ \{u_i(g(m/m_i'),t_i(m),\omega) - u_i(g(m/m_i'),t_i(m) + \varepsilon + \frac{\xi}{K},\omega)\}\\
&\leq \frac{v}{K} - (\varepsilon + \frac{\xi}{K})\rho < 0,
\end{aligned}$$

which are derived from the inequalities (1), (2), and (3).[13]

Next, suppose $i = 1$, and $m_j^k = \mu_j^k(\omega)$ for all $j \in N/\{1\}$. Then, agent 1 has incentive to announce m_1' instead of m_1, because $g(m)$ is independent of m_1^k, and

$$x_1(m/m_1') - x_1(m) = \frac{\xi}{K} > 0.$$

From the above arguments, we have proved that $\mu(\omega)$ is the unique iteratively honesty-proof message profile in $G^*(f,K)$.

Q.E.D.

Theorem 2 implies that every normative social choice function is implementable in iterative honesty-proofness. This is in contrast to the fact that any normative social choice function is never implementable in any purely self-interested solution concept such as Nash equilibrium.

Whether and whom the central planner will fine crucially depends on agent 1's $(K+1)-th$ opinion. Note that only the agent who tells a lie will be fined if and only if agent 1's $(K+1)-th$ opinion is honest. This implies that all agents will announce the honest message profile as the unique equilibrium even when agent 1 is not necessarily honesty-oriented but does dislike harming truth-telling agents.

[13] We must note that m_i' is never eliminated before m_i being eliminated.

4. Complete Information with No Fines

This section reconsiders the complete information environments and assumes $n=3$. In contrast to Section 3, we do *not* allow the central planner to fine agents, i.e., we assume $\varepsilon=0$ and $\xi=0$. Hence, we consider only mechanisms satisfying that

$$x_i(m)=0 \text{ for all } i\in N \text{ and all } m\in M .$$

We will simply write $u_i(a,\omega)$ instead of $u_i(a,0,\omega)$. We further confine our attention to mechanisms $\widetilde{G}=(M,g,x)$ where for every $i\in N$,

$$M_i=\Omega\times\{0,1,2,3\}.$$

Hence, each agent announces an integer in the set $\{0,1,2,3\}$ as well as a single element of Ω. For every $i\in N$, let $M_i=M_i^1\times M_i^2$ and $m_i=(m_i^1,m_i^2)$ where $M_i^1=\Omega$, $M_i^2=\{0,1,2,3\}$, and $m_i^k\in M_i^k$. Let $M^k=M_1^k\times M_2^k\times M_3^k$ and $m^k=(m_1^k,m_2^k,m_3^k)$ for each $k\in\{1,2\}$. The *honest message rule for each agent* i *at each state* $\omega\in\Omega$ is defined by

$$\widetilde{\mu}_i(\omega)=(\widetilde{\mu}_i^1(\omega),\widetilde{\mu}_i^2(\omega))=(\omega,0).$$

Hence, by announcing the honest message, each agent announces the integer 0 as her second opinion. Let $\widetilde{\mu}(\omega)=(\widetilde{\mu}_i(\omega))_{i\in N}$, $\widetilde{\mu}^1(\omega)=(\widetilde{\mu}_i^1(\omega))_{i\in N}$, and $\widetilde{\mu}^2(\omega)=(\widetilde{\mu}_i^2(\omega))_{i\in N}$.

4.1. Honesty-Proof Nash Equilibrium

This section assumes that agent 1 is honesty-oriented in a lexicographical sense that she prefers m_1 to m_1' if m_1 is the same as m_1' except for her first opinion, both m_1 and m_1' provide her with the same expected utility, and m_1 induces agent 1 to make the honest announcement as her first opinion, i.e.,

$$m_1^1=\widetilde{\mu}_1^1(\omega)\neq m_1'^1.$$

A Nash equilibrium message profile $m\in M$ in the game $(\widetilde{G},\omega)$ is said to be *honesty-proof* if either $m_1^1=\widetilde{\mu}_1^1(\omega)$, or

$$u_1(g(m),\omega)>u_1(g(m/m_1'),\omega),$$

where $m_1'=(\widetilde{\mu}_1^1(\omega),m_1^2)$. Hence, if m is a honesty-proof Nash equilibrium, then either agent 1 makes the honest announcement as her first opinion or the replacement of her first opinion by the honest announcement $m_1'=(\widetilde{\mu}_1^1(\omega),m_1^2)$ is not a best reply.

A social choice function $f\in F$ is said to be *implemented by the mechanism* $\widetilde{G}$ *in honesty-proof Nash equilibrium* if for every $\omega\in\Omega$, the honest message profile $\widetilde{\mu}(\omega)$ is the unique honesty-proof Nash equilibrium at the state ω in $\widetilde{G}$, and

$$g(\widetilde{\mu}(\omega))=f(\omega).$$

4.2. Specification of Mechanisms

For every $i \in N$, we define *agent* i*'s dictatorial social choice function* $d_i : \Omega \to A$ by

$$u_i(d_i(\omega), \omega) \geq u_i(a, \omega) \text{ for all } \omega \in \Omega \text{ and all } a \in A.$$

At every state $\omega \in \Omega$, agent i prefers $d_i(\omega)$ the best. We introduce the following two conditions on $(d_i)_{i \in N}$.

Condition 1: For every $\omega \in \Omega$, $\omega' \in \Omega$, every $i \in N$, and every $j \in N$,

$$u_i(d_i(\omega), \omega) > u_i(d_j(\omega'), \omega) \text{ if } d_i(\omega) \neq d_j(\omega').$$

The inequalities in Condition 1 imply that agent i strictly prefers her dictatorial choice $d_i(\omega)$ to any element in the range of agent $j's$ dictatorial social choice function.

Condition 2: For every $\omega \in \Omega$, and every $i \in N$, there exists $j \in N / \{i\}$ such that

$$d_j(\omega) \neq d_i(\omega).$$

We specify a *modulo mechanism* $\tilde{G} = \tilde{G}(f)$ as follows.[14] We define $\iota : M^2 \to N$ in ways that for every $m^2 \in M^2$, there exists $l \in \{-1, 0, 1, 2\}$ such that

$$m_1^2 + m_2^2 + m_3^2 = 3l + \iota(m^2).$$

The function ι defines the 'modulo game' where each agent $i \in N$ announces an integer m_i^2 in the set $\{0,1,2,3\}$ and agent $\iota(m^2)$ will be the winner. We specify g in ways that for every $m \in M$,

$$g(m) = d_i(m_i^1) \text{ if } i = \iota(m^2) \text{ and there exists no } \omega \in \Omega \text{ such that}$$
$$m_j = \tilde{\mu}_i(\omega) \text{ for two or more agents } j \in N,$$

and for every $\omega \in \Omega$,

$$g(m) = f(\omega) \text{ if } m_i = \tilde{\mu}_i(\omega) \text{ for two or more agents } i \in N.$$

The central planner chooses a pure alternative according to the simple lottery that the social choice function assigns to the majority opinion when it exists. When there exists no majority opinion, the winner i of the modulo game becomes dictatorial.

When all agents announce the honest message profile $\tilde{\mu}(\omega)$, the central planner chooses a pure alternative according to the simple lottery $f(\omega)$ that is assigned by the social choice function f to the state $\omega \in \Omega$, i.e., for every $\omega \in \Omega$,

$$g(\tilde{\mu}(\omega)) = f(\omega).$$

[14] See Maskin (1999) and Moore (1992). In a modulo mechanism, each agent announces an integer in a finite set as well as makes a single announcement about the state, and whenever agents make inconsistent announcements then they play the modulo game.

4.3. The Possibility Theorem

Note that for every $\omega \in \Omega$, and every $\omega' \in \Omega / \{\omega\}$, the message profile $\tilde{\mu}(\omega')$ is a Nash equilibrium in $(\tilde{G}(f), \omega)$, because for every $i \in N$, and every $m_i \in M_i$,

$$u_i(g(\mu(\omega')/m_i), \omega) = u_i(f(\omega'), \omega).$$

This implies that every inconstant social choice function $f \in F$ is never implemented by $\tilde{G}(f)$ in Nash equilibrium. In contrast, the following theorem states that with minor restrictions, every social choice function $f \in F$ can be implemented by $\tilde{G}(f)$ in honesty-proof Nash equilibrium. We introduce the following condition on the social choice function $f \in F$.

Condition 3: For every $\omega \in \Omega$, and every $\omega' \in \Omega / \{\omega\}$,

$$u_i(d_i(\omega), \omega) > u_i(f(\omega'), \omega) \text{ for two or more agents } i \in N.$$

Condition 3 implies that there exists no lottery in the range of the social choice function that is best preferred by two or more agents.

Theorem 3: *With Conditions 1 and 2, a social choice function* $f \in F$ *is implemented by* $\tilde{G}(f)$ *in honesty-proof Nash equilibrium if it satisfies Condition 3.*

Proof: Fix $\omega \in \Omega$ arbitrarily. Note that for every $i \in N$,

$$g(\tilde{\mu}(\omega)/m_i) = f(\omega) \text{ for all } m_i \in M_i.$$

Hence, $\tilde{\mu}(\omega)$ is an honesty-proof Nash equilibrium.

Fix $m \in M / \{\mu(\omega)\}$, and suppose that m is an honesty-proof Nash equilibrium. Suppose that $m = \tilde{\mu}(\omega')$ for some $\omega' \in \Omega / \{\omega\}$. Then, agent 1 has incentive to announce $\tilde{\mu}_1(\omega)$, because $g(\tilde{\mu}(\omega')/m_1) = f(\omega')$ for all $m_1 \in M_1$. This is a contradiction.

Suppose that there exist $\omega' \in \Omega$ and $i \in N$ such that $m_i \neq \tilde{\mu}_i(\omega')$, and $m_j = \tilde{\mu}_j(\omega')$ for all $j \in N / \{i\}$. Then, it follows that

$$g(m) = f_i(\omega').$$

From Condition 3, it follows that there exists $\tilde{i} \in N / \{i\}$ such that

$$u_{\tilde{i}}(d_{\tilde{i}}(\omega), \omega) > u_{\tilde{i}}(f(\omega'), \omega).$$

Note that there exists $m'_{\tilde{i}} \in M_{\tilde{i}} / \{m_{\tilde{i}}\}$ such that

$$m'^1_{\tilde{i}} = \omega, \quad m'^2_{\tilde{i}} \in \{1,2,3\}, \quad \tilde{i} = \iota(m^2 / m'^2_{\tilde{i}}),$$

and therefore,

$$g(m / m'_{\tilde{i}}) = d_{\tilde{i}}(\omega).$$

Hence, agent $\tilde{i}$ has incentive to announce $m'_{\tilde{i}}$ instead of $m_{\tilde{i}}$. This is a contradiction.

Suppose that there exists no $\omega' \in \Omega / \{\omega\}$ such that $m_i = \tilde{\mu}_i(\omega')$ for two or more agents $i \in N$. Then, it follows that

$$g(m) = d_{\iota(m^2)}(m^1_{\iota(m^2)}).$$

From Condition 1, it follows that if $d_{\iota(m^2)}(m^1_{\iota(m^2)}) \neq d_{\iota(m^2)}(\omega)$, then agent $\iota(m^2)$ has incentive to announce $m'_{\iota(m^2)}$ instead of $m_{\iota(m^2)}$, where $m'^1_{\iota(m^2)} = \omega$, $m'^2_{\iota(m^2)} \in \{1,2,3\}$, $\iota(m^2 / m'^2_{\iota(m^2)}) = \iota(m^2)$, and therefore,

$$g(m / m'_{\iota(m^2)}) = d_{\iota(m^2)}(\omega).$$

From Condition 2, it follows that if $d_{\iota(m^2)}(m^1_{\iota(m^2)}) = d_{\iota(m^2)}(\omega)$, then there exists $j \in N / \{\iota(m^2))\}$ who has incentive to announce m'_j instead of m_j, where $m'^1_j = \omega$, $m'^2_j \in \{1,2,3\}$, $\iota(m^2 / m'^2_j) = j$, and therefore,

$$g(m / m'_j) = d_j(\omega),$$

which is preferred to $g(m) = d_{\iota(m^2)}(\omega)$ by agent j. This is a contradiction.

From the above arguments, we have proved that $\tilde{\mu}(\omega)$ is the unique honesty-proof Nash equilibrium.

Q.E.D.

Theorem 3 implies that every normative social choice function is implementable in honesty-proof Nash equilibrium, where we use no fines. We must note that there may exist unwanted mixed strategy equilibria that may be consistent with the honesty-proofness. For example, consider a mixed message profile, according to which, for every $k \in \{1,2,3\}$, each agent $i \in N$ announces (ω, k) with probability $\frac{1}{3}$. The resultant lottery equals

$$\frac{d_1(\omega) + d_2(\omega) + d_3(\omega)}{3},$$

which may be different from the socially desired simple lottery $f(\omega)$ at the state ω. Note that this mixed message profile is a Nash equilibrium at the state ω. When each agent $i \in N$ deviates from this profile by announcing a message (ω', k), the resultant lottery equals

$$\frac{d_1(\omega) + d_2(\omega) + d_3(\omega)}{3} + \frac{d_i(\omega') - d_i(\omega)}{3},$$

which is not preferred to $\frac{d_1(\omega) + d_2(\omega) + d_3(\omega)}{3}$ by this agent. This mixed action profile is consistent with the honesty-proofness constraints, because it *does* induce every agent to announce the true state as her first opinion. Hence, it follows that when we take mixed strategy equilibria into account, the modulo mechanism may not be able to implement a social choice function even if we require the honesty-proofness in the above sense.

A possible way of solving this difficulty would be the following. Suppose that the integer 0 is regarded as being 'salient', and agent 1 is not only honesty-oriented but also *salience-oriented* in that at any state $\omega \in \Omega$, she surely announces the honest message $\tilde{\mu}_1(\omega) = (\omega, 0)$ whenever it is one of her best replies. Then, it follows that whenever all other agents play the mixed message profile above then agent 1 has incentive to announce the honest message $\tilde{\mu}_1(\omega)$ instead of the mixed message, because the resultant lottery equals $\frac{d_1(\omega) + d_2(\omega) + d_3(\omega)}{3}$, i.e., the honest message is one of her best replies. Hence, it follows that we can eliminate the mixed message profile when agent 1 is salience-oriented as well as honesty-oriented.

5. Incomplete Information: Lexicographical Approach

This section considers the incomplete information environments. Each agent $i \in N$ knows only her private signal $\omega_i \in \Omega_i$, where Ω_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., $\Omega \equiv \prod_{i \in N} \Omega_i$.[15] The private signal structure is given 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. For every $i \in N$, and every $j \in N / \{i\}$, let

$$p_{ij}(\omega_j \mid \omega_i) \equiv \sum_{\omega_{-i-j} \in \Omega_{-i-j}} p_i(\omega_{-i} \mid \omega_i),$$

and for every $(\omega_i, \omega_j) \in \Omega_i \times \Omega_j$ satisfying that $p_{ij}(\omega_j \mid \omega_i) \neq 0$, let

$$p_i(\omega_{-i-j} \mid \omega_i, \omega_j) \equiv \frac{p_i(\omega_{-i} \mid \omega_i)}{p_{ij}(\omega_j \mid \omega_i)},$$

and

$$p_i(\cdot \mid \omega_i, \omega_j) \equiv (p_i(\omega_{-i-j} \mid \omega_i, \omega_j))_{\omega_{-i-j} \in \Omega_{-i-j}} .$$

We assume that *three or more* agents are required to announce messages, i.e.,

$$n \geq 3 .$$

We assume that the central planner is allowed to fine agents where

$$\varepsilon > 0 \quad \text{and } \xi = 0 .$$[16]

We also assume that utilities are *quasi-linear* in that for every $i \in N$, and every $(a, t_i, \omega) \in A \times [-\varepsilon, 0] \times \Omega_i$,

$$u_i(a, t_i, \omega) = u_i(a, \omega) + t_i .$$

We consider only mechanisms $\hat{G} = (M, g, x)$ satisfying that for every $i \in N$,

$$M_i = \Omega_i^{K_i} .$$

Each agent $i \in N$ announces K_i elements of Ω_i at one time as her multiple opinions about her private signal. Let $M_i = M_i^1 \times \cdots \times M_i^{K_i}$ and $m_i = (m_i^1, ..., m_i^{K_i})$, where $M_i^k = \Omega_i$ and $m_i^k \in M_i^k$. A *message rule for each agent* $i \in N$ is defined by a function $\eta_i : \Omega_i \to M_i$. Let Ξ_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

[15] In the next section, we reconsider the incomplete information environments, where we do not assume the finiteness.

[16] This section requires a stronger version of honesty-proofness than the previous section in that every agent is honesty-oriented for *all* of her multiple announcements. This makes the fine ξ redundant, because even without this fine every agent may have incentive to make all of her announcements honest. On the other hand, when an agent is honesty-oriented only for her partial announcements, we may need a positive fine $\xi > 0$ in order for her to have strict incentive to make all of her announcements honest.

$$\Xi \equiv \prod_{i\in N} \Xi_i \text{ and } \eta(\omega) = (\eta_i(\omega_i))_{i\in N}.$$

The *honest message rule for each agent* $i \in N$ is defined by $\hat{\eta}_i = (\hat{\eta}_i^k)_{k=1}^{K_i} \in \Xi_i$ where

$$\hat{\eta}_i^k(\omega_i) = \omega_i \text{ for all } k \in \{1,...,K_i\} \text{ and all } \omega_i \in \Omega_i.$$

Let $\hat{\eta} = (\hat{\eta}_i)_{i\in N}$ denote the honest message rule profile.

A *mixed message for each agent* $i \in N$ is defined by a function $\lambda_i : M_i \to [0,1]$, where $\sum_{m_i\in M_i} \lambda_i(m_i) = 1$. A *mixed message rule for each agent* $i \in N$ is defined by a function $\chi_i : \Omega_i \to \Lambda_i$. A mixed message rule profile $\chi = (\chi_i)_{i\in N}$ is said to be a *mixed Bayesian Nash equilibrium in a mechanism* G if for every $i \in N$, every $\omega_i \in \Omega_i$,

$$E[\sum_{m\in M} \chi(\omega)(m)\{u_i(g(m),\omega) + x_i(m)\} \,|\, \omega_i]$$
$$\geq E[\sum_{m_{-i}\in M_{-i}} \chi_{-i}(\omega_{-i})(m_{-i})\{u_i(g(m/m_i'),\omega) + x_i(m/m_i')\} \,|\, \omega_i] \text{ for all } m_i' \in M_i,$$

where $\chi(\omega)(m) = \prod_{j\in N} \chi_j(\omega_j)(m_j)$, $\chi_{-i}(\omega_{-i})(m_{-i}) = \prod_{j\in N/\{i\}} \chi_j(\omega_j)(m_j)$, and $E[\cdot \,|\, \omega_i]$ implies the expected value conditional on ω_i. A social choice function $f \in F$ is said to be *implemented by a mechanism G in mixed Bayesian Nash equilibrium* if there exists a mixed Bayesian Nash equilibrium in G, and every mixed Bayesian Nash equilibrium χ in G always induces the socially desired simple lottery, i.e., for every $\omega \in \Omega$, and every $m \in M$,

$$g(m) = f(\omega) \text{ whenever } \chi(\omega)(m) > 0.$$

All agents are said to have *complete knowledge about their preferences* if a pair of distinct states $\omega \in \Omega$ and $\omega' \in \Omega/\{\omega\}$ is preference-equivalent whenever there exists at least one agent $i \in N$ such that $\omega_i = \omega_i'$. We can show that with complete knowledge and with a minor restriction on the class of mechanisms, no normative social choice function is implementable in mixed Bayesian Nash equilibrium. A mechanism G is said to be *regular* if for every $\omega \in \Omega$, there exists a mixed message profile $\lambda = (\lambda_i)_{i\in N}$ such that for every $i \in N$,

$$\sum_{m\in M} \lambda(m)\{u_i(g(m),\omega) + x_i(m)\}$$
$$\geq \sum_{m_{-i}\in M_{-i}} \lambda_{-i}(m_{-i})\{u_i(g(m/m_i'),\omega) + x_i(m/m_i')\} \text{ for all } m_i' \in M_i,$$

where $\lambda(m) = \prod_{j\in N} \lambda_j(m_j)$ and $\lambda_{-i}(m) = \prod_{j\in N/\{i\}} \lambda_j(m_j)$. Regularity requires that at every state $\omega \in \Omega$, there exists a mixed Nash equilibrium in the game with complete information defined by (G,ω).[17]

[17] Regularity is a condition regarding the plausibility of mechanisms, which is similar to, but not the same as, the regularity addressed in Abreu and Matsushima (1992b).

Proposition 4: *Suppose that all agents have complete knowledge about their preferences. Then, for every normative social choice function f, there exists no regular mechanism G that implements f in mixed Bayesian Nash equilibrium.*

Proof: Since each agent $i \in N$ has complete knowledge about her preference, it follows that without loss of generality, we can write $u_i(\cdot,\omega_i)$ instead of $u_i(\cdot,\omega)$ in this proof. Suppose that a regular mechanism G implements a social choice function f in mixed Bayesian Nash equilibrium. For every $\omega \in \Omega$, let λ^{ω} be a mixed Nash equilibrium message profile in the game (G,ω), satisfying that for every $i \in N$,

$$\sum_{m \in M} \lambda^{\omega}(m)\{u_i(g(m),\omega) + x_i(m)\}$$

$$\geq \sum_{m_{-i} \in M_{-i}} \lambda^{\omega}_{-i}(m_{-i})\{u_i(g(m/m_i'),\omega) + x_i(m/m_i')\} \text{ for all } m_i' \in M_i .$$

Without loss of generality, we assume that whenever ω and ω' are preference-equivalent then $\lambda^{\omega} = \lambda^{\omega'}$. Since all agents have complete knowledge about their preferences, it follows that there exists a mixed message rule profile $\chi = (\chi_i)_{i \in N}$ such that

$$\chi(\omega) = \lambda^{\omega} \text{ for all } \omega \in \Omega ,$$

and it is a mixed Bayesian Nash equilibrium in G. Note that whenever ω and ω' are preference-equivalent then $\chi(\omega)$ and $\chi(\omega')$ are the same. This is a contradiction when f is a normative social choice function.

Q.E.D.

In the same way as in Proposition 4, it follows that no normative social choice function is implementable in any purely self-interested solution concept in the incomplete information environments with complete knowledge and regularity. Moreover, it follows that no normative social choice function is virtually implementable in any purely self-interested solution concept in the incomplete information environments with complete knowledge and regularity, because any other social choice function that is sufficiently close to the normative social choice function must be normative.

A social choice functions f is said to be *incentive compatible* if for every $i \in N$, and every $\omega_i \in \Omega_i$,

$$E[u_i(f(\omega),\omega) \mid \omega_i] \geq E[u_i(f(\omega/\omega_i'),\omega) \mid \omega_i] \text{ for all } \omega_i' \in \Omega_i .$$

In other word, a social choice functions is incentive compatible if and only if the honest message rule profile is a Bayesian Nash equilibrium in the direct mechanism associated with this social choice function and with zero side payments.

5.1. Bayesian Iterative Honesty-Proofness

A message $m_i \in M_i$ for agent $i \in N$ is said to be *more honest than* a message $m_i' \in M_i / \{m_i\}$ when her private signal is given by $\omega_i \in \Omega_i$, if for every $k \in \{1,...,K_i\}$,

$$m_i'^k = \omega_i \text{ whenever } m_i^k = \omega_i,$$

and

$$\text{either } m_i'^k = \omega_i \text{ or } m_i'^k = m_i^k \text{ whenever } m_i^k \neq \omega_i.$$

We assume that every agent is honesty-oriented in a lexicographical sense that she prefers m_i' to m_i if m_i' is more honest than m_i and both m_i' and m_i provide her with the same expected utility. Hence, every agent is required to be honesty-oriented for all of her $K+1$ announcements. This is in contrast to Section 3 where only agent 1 is required to be honesty-oriented and she is so only for the $(K+1)-th$ announcement. We denote by $M_i(\omega_i, m_i) \subset M_i$ the set of all messages $m_i' \in M_i / \{m_i\}$ for agent i that are more honest than m_i when agent $i's$ private signal is given by $\omega_i \in \Omega_i$.

We introduce the solution concept named *Bayesian iterative honesty-proofness* as follows. For every $i \in N$, let $\Xi_i(1) \equiv \Xi_i$. For every integer $h \geq 2$, and every $i \in N$, let $\Xi_i(h)$ denote the set of all message rules $\eta_i \in \Xi_i(h-1)$ for agent i satisfying that there exist no $\eta_i' \in \Xi_i(h-1)$ and no $\omega_i \in \Omega_i$ such that

$$\eta_i'(\omega_i) \in M_i(\omega_i, \eta_i(\omega_i)),$$

and for every $\eta_{-i} \in \Xi_{-i}(h-1)$,

$$E[u_i(g(\eta(\omega)/\eta_i'(\omega_i)),\omega) + x_i(\eta(\omega)/\eta_i'(\omega_i)) \mid \omega_i]$$
$$\geq E[u_i(g(\eta(\omega)),\omega) + x_i(\eta(\omega)) \mid \omega_i],$$

where $\Xi(h-1) \equiv \prod_{i\in N} \Xi_i(h-1)$ and $\Xi_{-i}(h-1) \equiv \prod_{j\in N/\{i\}} \Xi_j(h-1)$. Hence, each agent never announces dominated messages with respect to the honesty-proofness in every round of iterative removal. Let $\Xi_i(\infty) \equiv \lim_{h\to\infty} \Xi_i(h)$ and $\Xi(\infty) \equiv \lim_{h\to\infty} \Xi(h)$. A message rule profile $\eta \in \Xi$ is said to *be Bayesian iteratively honesty-proof in the mechanism* $\hat{G}$ if $\eta \in \Xi(\infty)$. Note that the definition of Bayesian iterative honesty-proofness is irrelevant to the order of iterative removal. Note also that if there exists the unique Bayesian iteratively honesty-proof message rule profile $\eta \in \Xi$ in $\hat{G}$, then it is the unique mixed Bayesian Nash equilibrium χ in $\hat{G}$ satisfying that every agent $i \in N$ behaves as being honesty-oriented in the sense that for every $\omega_i \in \Omega$, every $m_i \in M_i$ satisfying $\chi_i(\omega_i)(m_i) > 0$, and every $m_i' \in M_i(\omega_i, m_i)$,

$$E[\sum_{m\in M} \chi(\omega)(m)\{u_i(g(m),\omega) + x_i(m)\} \mid \omega_i]$$
$$> E[\sum_{m_{-i}\in M_{-i}} \chi_{-i}(\omega_{-i})(m_{-i})\{u_i(g(m/m_i'),\omega) + x_i(m/m_i')\} \mid \omega_i].$$

A social choice function f is said to be *implemented by the mechanism* $\hat{G}$ *in Bayesian iterative honesty-proofness* if the honest message rule profile $\hat{\eta}$ is the only iteratively honesty-proof message rule profile in $\hat{G}$, and for every $\omega \in \Omega$,

$$g(\hat{\eta}(\omega)) = f(\omega),$$

and

$$x_i(\hat{\eta}(\omega)) = 0 \text{ for all } i \in N.$$

5.2. Specification of Mechanisms

We specify a mechanism $\hat{G} = \hat{G}(f, K)$ as follows. Let

$$K_i = K + 1 \text{ for all } i \in N.$$

For every $k \in \{1,...,K+1\}$, let $m^k = (m_1^k,...,m_n^k)$, $\eta^k = (\eta_i^k)_{i\in N}$, and $\eta^k(\omega) = (\eta_i^k(\omega_i))_{i\in N}$. We specify g by

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

Hence, for every $k \in \{1,...,K\}$, with probability $\frac{1}{K}$, the central planner chooses a pure alternative according to the lottery $f(m^k)$ that the social choice function f assigns to the profile of agents' k-th opinions m^k.

For every $i \in N$, and every $j \in N/\{i\}$, we define a function $s_{ij} : \prod_{i\in N} \Omega_i \to [-1,0]$ in ways that for every $\omega \in \Omega$,

$$s_{ij}(\omega) = \frac{-\{1 - p_i(\omega_{-i-j} \mid \omega_i, \omega_j)\}^2 - \sum_{\omega'_{-i-j} \in \Omega_{-i-j}/\{\omega_{-i-j}\}} p_i(\omega'_{-i-j} \mid \omega_i, \omega_j)^2}{2},$$

and for every $\omega \notin \Omega$,

$$s_{ij}(\omega) = -1.$$

For every $i \in N$, and every $m \in M$, we define $h(i,m)$ as the integer $k \in \{1,...,K\}$ satisfying that there exists $j \in N/\{i\}$ such that

$$m_j^k \neq m_j^{K+1},$$

and

$$m_{j'}^{k'} = m_{j'}^{K+1} \text{ for all } k' \in \{1,...,k-1\} \text{ and all } j' \in N/\{i\}.$$

If no such $k \in \{1,...,K\}$ exist, let $h(i,m) \equiv K+1$. For every $i \in N$, every $m \in M$, and every $k \in \{1,...,K\}$, we define $N(i,m,k) \subset N$ as the set of all agents $j \in N/\{i\}$ satisfying that $m_j^k \neq m_j^{K+1}$. Note that for every $i \in N$, and every $m \in M$, $N(i,m,K+1)$ is an empty set. For every $i \in N$, we specify x_i by

$$x_i(m) = \{\sum_{j\in N(i,m,k(i,m))} s_{ij}(m^{K+1}/m_i^{k(i,m)})\}\varepsilon.$$

Our construction is similar to Abreu and Matsushima (1992b) in that each agent makes announces multiple opinions about her private signal and the central planner fines the first deviants. Our construction, however, is simpler than Abreu and Matsushima, because we do not use the device of virtualness.[18] [19]

When all agents announce the honest message profile $\hat{\eta}(\omega)$, the central planner chooses a pure alternative according to the lottery $f(\omega)$ that is assigned by the social choice function f to the true state $\omega \in \Omega$ and no agents are fined, i.e., for every $\omega \in \Omega$,

$$g(\hat{\eta}(\omega)) = f(\omega),$$

and

$$x_i(\hat{\eta}(\omega)) = 0 \text{ for all } i \in N.$$

5.3. The Possibility Theorem

We introduce the following condition on the private signal structure p, which requires agents' private signals to be *correlated.*

Condition 4: For every $i \in N$, every $j \in N/\{i\}$, every $(\omega_i, \omega_j) \in \Omega_i \times \Omega_j$, and every $\omega_i' \in \Omega_i$,

$$p_i(\cdot | \omega_i, \omega_j) \neq p_i(\cdot | \omega_i', \omega_j) \text{ if } p_{ij}(\omega_j | \omega_i) \neq 0 \text{ and } p_{ij}(\omega_j | \omega_i') \neq 0.$$

Condition 4 holds generically in the set of possible private signal structures with correlations, but excludes the case that agents' private signals are independent.[20] The following theorem states that with Condition 4, every incentive compatible social choice function is implementable in Bayesian iterative honesty-proofness.

Theorem 5: *With Condition 4, for every incentive compatible social choice function f, there exists K such that f is implemented by the mechanism $\hat{G}(f, K)$ in Bayesian iterative honesty-proofness.*

Proof: Condition 4 implies that for every $i \in N$, every $j \in N/\{i\}$, every $(\omega_i, \omega_j) \in \Omega_i \times \Omega_j$ satisfying that $p_{ij}(\omega_j | \omega_i) \neq 0$, and every $\omega_i' \in \Omega/\{\omega_i\}$,

[18] Abreu and Matsushima (1992b) did not assume quasi-linearity.

[19] The mechanism, however, is complicated in that it depends on the functions (s_{ij}) defined by using the fine detail of the private signal structure p. In the next section, we will construct mechanisms that are much simpler in this respect.

[20] We must note that the genericity of the fact that each agent's private signal is well informative when agents' private signals are correlated crucially depends on what is the class of state spaces that we are considering. Neeman (1999) showed that the genericity of this informativeness does not hold when agents' preferences depend not only on their private signals but also on other factors. See, however, the next section of the present paper, where we will not require any restriction on the state space such as Condition 4.

$$E[s_{ij}(\omega) - s_{ij}(\omega / \omega_i') \,|\, \omega_i, \omega_j] > 0,$$

where $E[\cdot \,|\, \omega_i, \omega_j]$ implies the expectation value conditional on $(\omega_i, \omega_j) \in \Omega_i \times \Omega_j$. Hence, we can choose K so that for every $i \in N$, every $j \in N/\{i\}$, every $(\omega_i, \omega_j) \in \Omega_i \times \Omega_j$ satisfying that $p_{ij}(\omega_j \,|\, \omega_i) \neq 0$, and every $\omega_i' \in \Omega/\{\omega_i\}$,

$$\text{(4)} \qquad \frac{2v}{K} < E[s_i(\omega) - s_i(\omega / \omega_i') \,|\, \omega_i, \omega_j]\varepsilon .$$

For every $i \in N$, and every $\omega_i \in \Omega_i$, agent i always has incentive to announce $m_i^{K+1} = \hat{\eta}_i^{K+1}(\omega_i)$, because both $g(m)$ and $x_i(m)$ are independent of m_i^{K+1}. Hence, it follows that $\eta^{K+1} = \hat{\eta}^{K+1}$ for all $\eta \in \Xi(\infty)$.

Fix $\eta \in \Xi$ and $k \in \{1,...,K\}$ arbitrarily, and suppose that

$$\eta^{k'} = \hat{\eta}^{k'} \text{ for all } k' \in \{1,...,k-1\}.$$

Fix $i \in N$ and $\omega_i \in \Omega_i$ arbitrarily. Suppose $\eta_i^k(\omega_i) \neq \hat{\eta}_i^k(\omega_i)$. Let $\bar{\eta}_i \in \Xi_i$ be the message rule for agent i satisfying that $\bar{\eta}_i(\omega_i') = \eta_i(\omega_i')$ for all $\omega_i' \in \Omega_i/\{\omega_i\}$, $\bar{\eta}_i^h(\omega_i) = \eta_i^h(\omega_i)$ for all $h \neq k$, and $\bar{\eta}_i^k(\omega_i) = \hat{\eta}_i^k(\omega_i)$. Since $N(i, \eta(\omega), k) = N(i, \eta(\omega)/\bar{\eta}_i(\omega_i), k)$, it follows that

$$\begin{aligned}
&E[u_i(g(\eta(\omega)), \omega) + x_i(\eta(\omega)) \,|\, \omega_i] \\
&- E[u_i(g(\eta(\omega)/\bar{\eta}_i(\omega_i)), \omega) + x_i(\eta(\omega)/\bar{\eta}_i(\omega_i)) \,|\, \omega_i] \\
&= \frac{1}{K} E[u_i(f(\eta^k(\omega)), \omega) - u_i(f(\eta^k(\omega)/\hat{\eta}_i^k(\omega_i)), \omega) \,|\, \omega_i] \\
&+ E[\sum_{j \in N(i, \eta(\omega), k)} \{s_{ij}(\omega / \eta_i^k(\omega_i)) - s_{ij}(\omega)\} \,|\, \omega_i]\varepsilon .
\end{aligned}$$

If $N(i, \eta(\omega), k)$ is empty for all $\omega_{-i} \in \Omega_{-i}$, then it follows that $\eta_{-i}^k = \hat{\eta}_{-i}^k$, and therefore, the utility difference above equals

$$\frac{1}{K} E[u_i(f(\omega / \eta_i^k(\omega_i)), \omega) - u_i(f(\omega), \omega) \,|\, \omega_i],$$

which is less than or equals zero, because f is incentive compatible. If there exist $j \in N/\{i\}$ and $\omega_j \in \Omega_j$ such that $j \in N(i, \eta(\omega), k)$ for some $\omega_{-i-j} \in \Omega_{-i-j}$, then it follows that $j \in N(i, \eta(\omega), k)$ for all $\omega_{-i-j} \in \Omega_{-i-j}$, and therefore, the utility difference above is less than or equals

$$\begin{aligned}
&\frac{1}{K} E[u_i(f(\omega / \eta_i^k(\omega_i)), \omega) - u_i(f(\omega), \omega) \,|\, \omega_i] + \frac{2v}{K} \sum_{\omega_{-i} \in \Omega_{-i} : \eta_{-i}(\omega_{-i}) \neq \omega_{-i}} p_i(\omega_{-i} \,|\, \omega_i) \\
&+ E[\sum_{j \in N(i, \eta(\omega), k)} \{s_{ij}(\omega / \eta_i^k(\omega_i)) - s_{ij}(\omega)\} \,|\, \omega_i]\varepsilon \\
&\leq \frac{1}{K} E[u_i(f(\omega / \eta_i^k(\omega_i)), \omega) - u_i(f(\omega), \omega) \,|\, \omega_i] \\
&+ \frac{2v}{K} \sum_{\omega_{-i} \in \Omega_{-i}} |N(i, \eta(\omega), k)| p_i(\omega_{-i} \,|\, \omega_i) + E[\sum_{j \in N(i, \eta(\omega), k)} \{s_{ij}(\omega / \eta_i^k(\omega_i)) - s_{ij}(\omega)\} \,|\, \omega_i]\varepsilon
\end{aligned}$$

$$= \frac{1}{K} E[u_i(f(\omega / \eta_i^k(\omega_i)), \omega) - u_i(f(\omega), \omega) \,|\, \omega_i]$$

$$+ \sum_{\substack{j \in N / \{i\}, \omega_j \in \Omega_j : \\ \eta_j^k(\omega_j) \neq \omega_j}} \{ \frac{2v}{K} - E[s_{ij}(\omega) - s_{ij}(\omega / \eta_i^k(\omega_i)) \,|\, \omega_i, \omega_j]\varepsilon \} p_{ij}(\omega_j \,|\, \omega_i),$$

which is negative, because f is incentive compatible and the inequalities (4) hold. Hence, it follows that $\eta^k = \hat{\eta}^k$ for all $\eta \in \Xi(\infty)$, and therefore, we have proved that $\hat{\eta}$ is the unique Bayesian honesty-proof message rule profile in $\hat{G}(f, K)$.

Q.E.D.

6. Incomplete Information: Positive Cost Approach

This section reconsiders the incomplete information environments. We assume that utilities are quasi-linear and the central planner is allowed to fine agents, where $\varepsilon > 0$, $\xi = 0$, and ε may be close to zero. In contrast to Section 5, we assume that it is *costly* for every agent to lie as her $K+1$ announcement. The cost of lying can be as close to zero as possible. Based on this positive cost hypothesis, together with agents' lexicographical preferences on honest reporting, we will introduce a modified version of Bayesian iterative honesty-proofness named Bayesian iterative honesty-proofness with positive cost, and show that every social choice function is implementable in this solution concept. This result is very permissive. We do not require any restriction on the private signal structure such as Condition 4. Hence, we allow agents' private signals to be independent. We do not assume that the set of state is finite. We do not assume that the number of agents is three or more, i.e., we assume only $n \geq 2$. Of particular importance, the constructed mechanism will be universal in the sense that it does not depend on the private signal structure $p = (p_i(\cdot \mid \omega_i))_{i \in N, \omega_i \in \Omega_i}$.

6.1. Specification of Mechanisms

We specify a mechanism $\hat{G}^* = \hat{G}^*(f, K)$ as follows. For every $i \in N$,

$$K_i = K+1 \text{ for all } i \in N.$$

For every $m \in M$, and every $a \in A$,

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

For every $i \in N$, and $m \in M$,

$$x_i(m) = -\hat{\varepsilon} \text{ if } m_i^{h(i,m)} \neq m_i^{K+1},$$

and

$$x_i(m) = 0 \text{ if } m_i^{h(i,m)} = m_i^{K+1},$$

where $\hat{\varepsilon}$ is a positive real number satisfying

$$0 < \hat{\varepsilon} \leq \varepsilon.$$

Note that when all agents announce the honest message profile $\hat{\eta}(\omega)$, the central planner chooses a pure alternative according to the lottery $f(\omega)$ that is assigned by the social choice function f to the true state $\omega \in \Omega$ and no agents are fined, i.e., for every $\omega \in \Omega$,

$$g(\hat{\eta}(\omega)) = f(\omega),$$

and

$$x_i(\hat{\eta}(\omega)) = 0 \text{ for all } i \in N.$$

Choose a positive real number $v > 0$ satisfying that

$$v \geq \left| u_i(a,\omega) - u_i(a',\omega) \right| \text{ for all } i \in N, \text{ all } \omega \in \Omega, \text{ all } a \in A, \text{ and all } a' \in A,$$

where v is regarded as the upper bound of agents' utility differences. We specify K as a positive integer satisfying that

(5) $$\hat{\varepsilon} \geq \frac{2v}{K}.$$

6.2. Bayesian Iterative Honesty-Proofness with Positive Cost

We assume that whenever each agent $i \in N$ announces a message $m_i \in M$ whose $(K+1)-th$ opinion is dishonest, i.e., $m_i^{K+1} \neq \hat{\eta}_i^{K+1}(\omega)$, then she will experience discomfort and incur a positive disutility $c > 0$. Moreover, every agent is honesty-oriented for the other opinions than her $(K+1)-th$ opinion in the lexicographical sense like Section 5. Hence, the total cost of lying is at most c irrespective of K. Hence, by letting c close to zero, we can make the upper bound of the total cost of lying in any mechanism as small as possible

Based on this, we define the solution concept named Bayesian iterative honesty-proofness with positive cost c as follows. For every $i \in N$, let $\Xi_i^*(1)$ denote the set of all message rules $\eta_i \in \Xi_i$ for agent i satisfying that for every $\omega_i \in \Omega_i$, either

$$\eta_i^{K+1}(\omega_i) = \hat{\eta}_i^{K+1}(\omega_i),$$

or there exists $\eta_{-i} \in \Xi_{-i}$ such that

$$E[u_i(g(\eta(\omega)), x_i(\eta(\omega)), \omega) \mid \omega_i] - c$$
$$\geq E[u_i(g(\eta(\omega)/m_i), x_i(\eta(\omega)/m_i), \omega) \mid \omega_i],$$

where $m_i = (\eta_i^1(\omega_i), ..., \eta_i^K(\omega_i), \hat{\eta}_i^{K+1}(\omega_i))$. This first round of iterative removal implies that each agent $i \in N$ is honesty-oriented in the sense that for every $m \in M$, whenever $m_i^{K+1} \neq \mu_i^{K+1}(\omega)$ and m_i' provides her with at least the expected utility induced by m_i minus the cost of lying c, then she never announces m_i. For every integer $h \geq 2$, and every $i \in N$, let $\Xi_i^*(h)$ denote the set of all message rules $\eta_i \in \Xi_i(h-1)$ for agent i satisfying that there exist no $\eta_i' \in \Xi_i(h-1)$ and no $\omega_i \in \Omega_i$ such that

$$\eta_i'(\omega_i) \in M_i(\omega_i, \eta_i(\omega_i)),$$

and for every $\eta_{-i} \in \Xi_{-i}(h-1)$,

$$E[u_i(g(\eta(\omega)/\eta_i'(\omega_i)), \omega) + x_i(\eta(\omega)/\eta_i'(\omega_i)) \mid \omega_i]$$
$$\geq E[u_i(g(\eta(\omega)), \omega) + x_i(\eta(\omega)) \mid \omega_i],$$

where $\Xi^*(h-1) \equiv \prod_{i \in N} \Xi_i^*(h-1)$ and $\Xi_{-i}^*(h-1) \equiv \prod_{j \in N/\{i\}} \Xi_j^*(h-1)$. Hence, in every round of iterative removal except for the first round, each agent is honesty-oriented only in the lexicographical sense. Let $\Xi_i^*(\infty) \equiv \lim_{h \to \infty} \Xi_i^*(h)$ and $\Xi^*(\infty) \equiv \lim_{h \to \infty} \Xi^*(h)$. A message rule profile $\eta \in \Xi$ is said to *be Bayesian iteratively honesty-proof with positive cost c in the* *mechanism* $\hat{G}^*(f,K)$ if $\eta\in\Xi^*(\infty)$. Note that the definition of Bayesian iterative honesty-proofness with positive cost is irrelevant to the order of iterative removal. Note also that if there exists the unique Bayesian iteratively honesty-proof message rule profile with positive cost c in $\hat{G}$, then it is the unique mixed Bayesian Nash equilibrium with positive cost c in $\hat{G}$ on the assumption that each agent's telling a lie as her $(K+1)-th$ announcement incurs a positive cost c. A social choice function f is said to be *implemented by the mechanism* $\hat{G}^*(f,K)$ *in Bayesian iterative honesty-proofness with positive cost* c if the honest message rule profile $\hat{\eta}$ is the only iteratively honesty-proof message rule profile in $\hat{G}^*(f,K)$, and for every $\omega\in\Omega$,

$$g(\hat{\eta}(\omega))=f(\omega),$$

and

$$x_i(\hat{\eta}(\omega))=0 \text{ for all } i\in N.$$

6.3. The Possibility Theorem

We use only a smaller fine than the cost of lying, i.e.,

(6) $$c>\hat{\varepsilon}>0.$$

The following theorem states that every incentive compatible social choice function is implementable in Bayesian iterative honesty-proofness with positive cost c.

Theorem 6: *Every incentive compatible social choice function* f *is implemented by the mechanism* $\hat{G}^*(f,K)$ *in Bayesian iterative honesty-proofness with positive cost* c.

Proof: For every $i\in N$, and every $\omega_i\in\Omega_i$, agent i always has incentive to announce $m_i^{K+1}=\hat{\eta}_i^{K+1}(\omega_i)$, because $g(m)$ is independent of m_i^{K+1} and the inequality (6) implies that $x_i(m)-x_i(m')$ is less than the cost of lying c. Hence, it follows that $\eta^{K+1}=\hat{\eta}^{K+1}$ for all $\eta\in\Xi^*(\infty)$.

Fix $\eta\in\Xi$ and $k\in\{1,...,K\}$ arbitrarily, and suppose that

$$\eta^{k'}=\hat{\eta}^{k'} \text{ for all } k'\in\{1,...,k-1,K\}.$$

Fix $i\in N$ and $\omega_i\in\Omega_i$ arbitrarily. Suppose $\eta_i^k(\omega_i)\neq\hat{\eta}_i^k(\omega_i)$. Let $\overline{\eta}_i\in\Xi_i$ be the message rule for agent i satisfying that $\overline{\eta}_i(\omega_i')=\eta_i(\omega_i')$ for all $\omega_i'\in\Omega_i/\{\omega_i\}$, $\overline{\eta}_i^h(\omega_i)=\eta_i^h(\omega_i)$ for all $h\neq k$, and $\overline{\eta}_i^k(\omega_i)=\hat{\eta}_i^k(\omega_i)$. Note from the inequality (5) and the incentive compatibility that

$$E[u_i(g(\eta(\omega)),\omega)+x_i(\eta(\omega))\,|\,\omega_i]$$
$$-E[u_i(g(\eta(\omega)/\overline{\eta}_i(\omega_i)),\omega)+x_i(\eta(\omega)/\overline{\eta}_i(\omega_i))\,|\,\omega_i]$$

$$
\begin{aligned}
&\leq \frac{1}{K} E[u_i(f(\eta^k(\omega)),\omega) \\
&+ u_i(f(\eta^k(\omega)/\hat{\eta}_i^k(\omega_i)),\omega) \mid \omega_i, \eta_{-i}^k(\omega) \neq \hat{\eta}_{-i}^k(\omega)]\; p_i(\eta_{-i}^k(\omega) \neq \hat{\eta}_{-i}^k(\omega) \mid \omega_i) \\
&- \frac{2v}{K} p_i(\eta_{-i}^k(\omega) \neq \hat{\eta}_{-i}^k(\omega) \mid \omega_i) + \frac{1}{K} E[u_i(f(\eta^k(\omega)),\omega) \\
&+ u_i(f(\eta^k(\omega)/\hat{\eta}_i^k(\omega_i)),\omega) \mid \omega_i, \eta_{-i}^k(\omega) = \hat{\eta}_{-i}^k(\omega)]\; p_i(\eta_{-i}^k(\omega) = \hat{\eta}_{-i}^k(\omega) \mid \omega_i) \\
&= \frac{1}{K} E[u_i(f(\hat{\eta}^k(\omega)/\eta_i^k(\omega)),\omega) + u_i(f(\hat{\eta}^k(\omega)),\omega) \mid \omega_i] \\
&\leq 0,
\end{aligned}
$$

where $p_i(\eta_{-i}^k(\omega) = \hat{\eta}_{-i}^k(\omega) \mid \omega_i)$ is the probability conditional on ω_i that the private signal profile $\omega \in \Omega$ satisfies $\eta_{-i}^k(\omega) = \hat{\eta}_{-i}^k(\omega)$, $E[\cdot \mid \omega_i, \eta_{-i}^k(\omega) = \hat{\eta}_{-i}^k(\omega)]$ implies the expected value conditional on ω_i and on the fact that $p_i(\eta_{-i}^k(\omega) = \hat{\eta}_{-i}^k(\omega) \mid \omega_i)$, and similarly, $p_i(\eta_{-i}^k(\omega) \neq \hat{\eta}_{-i}^k(\omega) \mid \omega_i)$ and $E[\cdot \mid \omega_i, \eta_{-i}^k(\omega) \neq \hat{\eta}_{-i}^k(\omega)]$ are defined. Hence, it follows that $\eta^k = \hat{\eta}^k$ for all $\eta \in \Xi^*(\infty)$, and therefore, we have proved that $\hat{\eta}$ is the unique Bayesian honesty-proof message rule profile with positive cost c in $\hat{G}^*(f,K)$.

Q.E.D.

Theorem 6 implies that whenever the cost of lying c is positive then every incentive compatible social choice function is implementable in Bayesian iterative honesty-proofness with positive cost c. Note that the specification of the mechanism $\hat{G}^* = \hat{G}^*(f,K)$ depends only on the cost of lying c, the social choice function f, and the upper bound of agents' utility differences v. Hence, given (c,f,v), the central planner can construct the mechanism without any knowledge about the private signal structure $p = (p_i(\cdot \mid \omega_i))_{i \in N, \omega_i \in \Omega_i}$. In this sense the mechanism is 'universal'. This point is in contrast to the standard model with incomplete information where the constructed mechanisms crucially depend on the fine detail of the private signal structure, and therefore, it would be difficult to use in practice.

Moreover, suppose that truth telling is an *ex post* equilibrium in the revelation game in that for every $i \in N$, and every $\omega \in \Omega$,

$$u_i(f(\omega),\omega) \geq u_i(f(\omega/\omega_i'),\omega) \text{ for all } \omega_i' \in \Omega_i.$$

Then, it follows that irrespective of the private signal structure $p = (p_i(\cdot \mid \omega_i))_{i \in N, \omega_i \in \Omega_i}$, the honest message rule profile $\hat{\eta}$ is always the unique Bayesian iteratively honesty-proof message rule profile in the mechanism $\hat{G}^*(f,K)$. This implies that even agents do not need to know what is the correct private signal structure $p = (p_i(\cdot \mid \omega_i))_{i \in N, \omega_i \in \Omega_i}$.[21]

[21] For characterization of ex-post equilibrium concept, see Chung and Ely (2002) and Bergemann and Morris (2002).

7. Concluding Remarks

This paper investigated implementation of social choice functions when agents are honesty-oriented only in marginal ways. In the complete information environments with small fines, it was shown that every social choice function is implementable when there exists a single agent who is not only purely self-interested but also honesty-oriented in a lexicographical way. Hence, all social choice functions that are important but never implementable in the standard implementation model, such as normative social choice functions, are implementable without contradicting the purely self-interested motive. This result could be extended to the incomplete information environments with quasi-linearity and with correlated private signals. Next, we assumed that it is costly for each agent to report dishonestly, but this cost can be as small as possible. It was shown that without any restriction on the model structure, every incentive compatible social choice function could be implemented by the mechanism that is universal in the sense that it does not depend on the private signal structure.

An underlying assumption in this paper is that agents either play a one-shot game or behave myopically in a repeated situation. When an agent is forward-looking and the cost of lying are sufficiently small, it might be the case that she is willing to manipulate information in order to convince the others that she is not honesty-oriented and likes harming others whose reports are against her taste. An important question in future researches might be when and how we can construct a mechanism in which honest behavior is endogenously stabilized in the dynamic process of evolution and learning.[22] This, however, is beyond the purpose of the paper.

[22] There exist recent works on evolution and learning in the implementation problem such as Cabrales (1999) and Matsushima (2002), which assumed naïve adaptive dynamics without forward looking players. Moreover, Kandori (2002) applied stochastic evolution to psychological games a la Rabin (1993).

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 (1999): "Moral Hazard, Insurance, and Some Collusion," forthcoming in *Journal of Economic Behavior and Organization*.

Bergemann, D. and S. Morris (2002): "Robust Mechanism Design," mimeo, Yale University.

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

Chung, Kim-Sau and J. Ely (2002): "Ex-Post Incentive Compatible Mechanism Design," mimeo, Northwestern University.

D'Aspremont, C. and L. Gevers (1977): "Equity and the Informational Basis of Collective Choice," *Review of Economic Studies* 46, 199-210.

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

Deschamps, R. and L. Gevers (1978): "Leximin and Utilitarian Rules: A Joint Characterization," *Journal of Economic Theory* 17, 143-163.

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

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.

Fehr, E. and K. Schmidt (2001): "A Theory of Fairness and Reciprocity: Evidence and Economic Applications," Working Paper No. 75, University of Zurich.

Geanakoplos, J., D. Pearce, and E. Stacchetti (1989): "Psychological Games and Sequential Rationality," *Games and Economic Behavior* 1, 60-79.

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

Kandori, M. (2002): "The Erosion and Sustainability of Norms and Morale," Discussion Paper CIRJE-F-169, University of Tokyo.

Krishna, V. (2002): *Auction Theory*, Academic Press.

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

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.

Matsushima, H. (2001): “Stability and Implementation via Simple Mechanisms in the Complete Information Environments,” Discussion Paper CIRJE-F-147, Faculty of Economics, University of Tokyo.

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.

Moore, J. and R. Repullo (1988): “Subgame Perfect Implementation,” *Econometrica* 46, 1191-1220.

Neeman, Z. (1999): The Relevance of Private Information in Mechanism Design,” mimeo.

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.

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

Rabin, M. (1993): “Incorporating Fairness into Game Theory,” *American Economic Review* 83, 1281-1320.

Rabin, M. (1998): “Psychology and Economics,” *Journal of Economic Literature* 36, 11-46.

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.