Role of Honesty in Full Implementation [+]

Hitoshi Matsushima[*]

Faculty of Economics, University of Tokyo

March 17, 2006

(First Version: March 4, 2002)

[+] This paper is a revised version of the manuscript entitled "Non-Consequential Moral Preferences, Detail-Free Implementation, and Representative Systems" (Discussion Paper CIRJE-F-304, Faculty of Economics, University of Tokyo, 2004). The research for this paper was supported by a Grant-In-Aid for Scientific Research (KAKENHI 15330036) from the Japan Society for the Promotion of Science (JSPS) and the Ministry of Education, Culture, Sports, Science and Technology (MEXT) of the Japanese government and a grant from the Center for Advanced Research in Finance (CARF) at the University of Tokyo.

[*] Faculty of Economics, University of Tokyo, Hongo, Bunkyo-ku, Tokyo 113-0033, Japan. E-mail: hitoshi@e.u-tokyo.ac.jp

Abstract

This paper introduces a new concept for full implementation that takes into account agents' preferences for understanding how the "process" works. We assume that the agents have an intrinsic preference for honesty in the sense that they dislike the idea of lying when it does not influence their welfare but instead goes against the intention of the principal. We show that the presence of such preferences functions very effectively in eliminating unwanted equilibria from the practical perspectives, even if the degree of preference for honesty is small. The mechanisms designed are detail-free and involve only small fines.

Keywords: Preferences for Honesty, Detail-free, Full Implementation, Small Fines, Permissive Results.

JEL Classification Numbers: C72, D71, D78, H41

1. Introduction

This paper introduces a new concept for *full* implementation that takes into account agents' preferences for understanding (not just the consequence but also) *how the "process" works*. We investigate environments in which a principal is unaware of the desirable alternative to be chosen, even though there exist multiple agents and they do have information about such an alternative. The principal delegates the alternative choice to these agents by designing an appropriate mechanism, according to which each agent is required to make *multiple* announcements about this alternative. The crucial assumption of this paper is that each of these agents has an *intrinsic preference for honesty* in the sense that she dislikes the idea of telling "white lies" that do not influence her welfare but instead go against the intention of the principal. This paper shows that with this assumption, it is very easy for the latter to incentivize the agents into telling the truth as the *unique* iteratively undominated strategy profile, thereby implementing the desirable alternative fully, and exactly, even if she does not have any information about this alternative in advance.

First, we consider a situation in which there exist *three* agents who have a *full* knowledge of the desirable alternative. Using only small fines, we show a very permissive result that there exists an *almighty* mechanism according to which the principal can achieve *any* alternative, provided that these agents regard this alternative as being desirable. This mechanism is *detail-free* in a very strict sense, i.e., it does not depend on the details of model specifications such as the *state space*. In this mechanism, the alternative choice and

the monetary transfer to an agent are never influenced by her first announcement. Since this agent has an intrinsic preference for honesty, it follows that she has a strict incentive to make her first announcement honestly. By using this honest announcement as a *reference*, and by punishing any other agent who is the first to deviate from this reference, the principal can incentivize all agents into telling the truth as the unique iteratively undominated strategy profile.

Second, we consider a Bayesian environment in which there exist multiple agents who do not have a full knowledge but have their respective *private* signals concerning the desirable alternative. We cannot use another agent's announcement as a reference to determine whether an agent is telling the truth because the private signals of all agents are different from each other. However, we can show that whenever the monetary fine is close to zero, the principal can incentivize each agent into telling the truth by regarding her *own* first announcement as the reference. Based on Abreu and Matsushima (1992b), without contradicting the requirement of full implementation, we can make monetary transfers as close to zero as possible. Hence, it follows from these observations that we can obtain a very permissive result that *every* incentive compatible social choice function is fully implementable in iterative dominance. We do not need any other condition apart from incentive compatibility. The designed mechanism is detail-free, i.e., it does not depend on the probability function on the state space. These features are in contrast with the previous works in the implementation literature, where agents' intrinsic preferences for honesty were not generally taken into account.[1]

[1] See survey articles such as Moore (1992), Palfrey (1992), Osborne and Rubinstein (1994, Chapter 10), and Maskin

The class of preferences for honesty considered in this paper covers many cases of the previous models such as Baiman and Lewis (1989) and Gneezy (2005). Baiman and Lewis investigated the threshold model in which each individual experiences fixed disutility from lying. Gneezy conducted laboratory experiments and showed that whether or not an individual lies depends on not just her own payoff but also other persons' payoffs. Our permissive results are independent of the degree to which the agents prefer to be honest. In this respect, we can state that our results hold even under the *minimal* requirement of the agents' preferences for honesty.

This paper is organized as follows. Section 2 shows the basic result of full implementation where the three agents have a full knowledge of the desirable alternative. Section 3 investigates the Bayesian environments and shows that incentive compatibility is necessary and sufficient for full implementation.

---

and Sjöström (2002). However, there are certain exceptions such as Glazer and Rubinstein (1998) and Eliaz (2002).

2. Basic Results

Consider a situation where a principal is unable to choose the desirable alternative from a nonempty and finite set of alternatives A. Further, there exist three agents, i.e., agents 1, 2, and 3, who have a full knowledge of the type of alternative that the principal should choose. The latter delegates the alternative choice to the three agents using the following process, which is a simpler version of the Abreu-Matsushima mechanism (Abreu and Matsushima (1992a, 1994)). The principal requires each agent to announce K number of times the choice that she should make, following which she randomly selects one announcement profile from the first $K-1$ profiles. Here, $K>0$ is a *sufficiently large* positive integer. If at least two agents announce the same alternative, she chooses that alternative. In the absence of such an alternative, she chooses the "status quo" that is given by $\bar{a} \in A$. She imposes a fine of $\varepsilon > 0$ if and only if the agent is either agent 2 or agent 3 and is the first to deviate from the first announcement made by agent 1. We assume that the monetary fine ε is close to zero.

Formally, we specify a mechanism $G=(M,g,t)$ as follows, where M_i is the set of *messages* for each agent i, $m_i \in M_i$, $M = \prod_{i\in\{1,2,3\}} M_i$, Δ denotes the set of lotteries over alternatives, $x: M \to \Delta$, $t=(t_i)_{i\in N}$, and $t_i : M \to \{-\varepsilon, 0\}$. When the agents announce a message profile $m=(m_i)_{i\in N} \in M$, the principal chooses any alternative $a \in A$ with the probability $x(m)[a]$ and makes a monetary transfer $t_i(m)$ to each agent i with certainty. Let

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

Let $M_i = M_{i,1} \times \cdots \times M_{i,K}$, where $M_{i,k} = A$ for all $k \in \{1,...,K\}$. For every $m \in M$, let

$$x(m)[a] = \frac{\#\{k \in \{2,...,K\} \mid m_{i,k} = a \text{ for two or three agents}\}}{K-1} \text{ for all } a \neq \bar{a},$$

$$x(m)[\bar{a}] = 1 - \sum_{a \neq \bar{a}} x(m)[a],$$

$$t_1(m) = 0,$$

for every $i \in \{2,3\}$,

$$t_i(m) = -\varepsilon \qquad \text{if there exist } k \in \{1,...,K\} \text{ such that } m_{i,k} \neq m_{1,1} \text{ and}$$

$$m_{2,h} = m_{3,h} = m_{1,1} \text{ for all } h \in \{1,...,k-1\},$$

and

$$t_i(m) = 0 \qquad \text{if there exists no such } k .$$

For every $k \in \{2,...,K\}$, the principal selects the $k-th$ announcement profile $(m_{1,k}, m_{2,k}, m_{3,k}) \in A^3$ with probability $\frac{1}{K-1}$ and chooses any alternative $a \in A$ if at least two agents announce this alternative, i.e., $m_{i,k} = a$ for at least two agents $i \in \{1,2,3\}$. In the absence of such an alternative, she chooses the status quo $\bar{a}$. Each agent $i \in \{2,3\}$ is fined if and only if she is the first to deviate from the first announcement $m_{1,1}$ made by agent 1. It should be noted that the latter is never fined.

Choose an arbitrary alternative $a^* \in A$ and regard it as the *desirable* alternative that the principal should choose. A preference for each agent $i \in \{1,2,3\}$, denoted by $\succ_{\sim i}$, is

defined on the set $\Delta \times \{-\varepsilon, 0\} \times M$. Here, $(\alpha, r_i, m) \underset{\sim}{\succ}_i (\alpha', r_i', m')$ implies "agent i does not prefer (α', r_i', m') to (α, r_i, m) ". $(\alpha, r_i, m) \succ_i (\alpha', r_i', m')$ implies " $(\alpha, r_i, m) \underset{\sim}{\succ}_i (\alpha', r_i', m')$ but $\sim[(\alpha', r_i', m') \underset{\sim}{\succ}_i (\alpha, r_i, m)]$; in other words, agent i prefers (α, r_i, m) to (α', r_i', m') ". Moreover, $(\alpha, r_i, m) \sim_i (\alpha', r_i', m')$ implies " $(\alpha, r_i, m) \underset{\sim}{\succ}_i (\alpha', r_i', m')$ and $(\alpha', r_i', m') \underset{\sim}{\succ}_i (\alpha, r_i, m)$; in other words, agent i is indifferent between (α, r_i, m) and (α', r_i', m') ". Let $\underset{\sim}{\succ} = (\underset{\sim}{\succ}_i)_{i \in \{1,2,3\}}$.

Condition 1: For every $i \in \{1,2,3\}$, $(\alpha, r_i, m) \in \Delta \times \{-\varepsilon, 0\} \times M$, and $m_i' \in M_i \setminus \{m_i\}$,

(1) $(\alpha, r_i, (m_i', m_{-i})) \succ_i (\alpha, r_i, m)$ if $m_{i,k}' \in \{a^*, m_{i,k}\}$ and $[m_{i,k} = a^*] \Rightarrow [m_{i,k}' = a^*]$ for all $k \in \{1, ..., K\}$.

Condition 1 implies that each agent is sufficiently honest to dislike any *white lie* that does not ever influence the alternative choice and the monetary transfer made to her.

Condition 2: For every $i \in \{1,2,3\}$, $(\alpha, m) \in \Delta \times M$, and $\alpha' \in \Delta$,

(2) $(\alpha', 0, m) \succ_i (\alpha, -\varepsilon, m)$ if $\max_{a \in A} |\alpha'[a] - \alpha[a]| \leq \dfrac{1}{K-1}$.

Condition 2 along with a sufficiently large K implies that the utility difference between two lotteries is almost negligible if these lotteries are close to each other in terms of the amount.

A combination $(G,\underset{\sim}{\succ})$ defines a *game*. The solution concept is *iterative dominance.* Let $M_i^{(0)} = M_i$ and $M^{(0)} = \prod_{i\in\{1,2,3\}} M_i^{(0)}$. Recursively, for every $\lambda = 1,2,...$, let $M_i^{(\lambda)}$ denote the set of messages $m_i \in M_i^{(\lambda-1)}$ for each agent i that are *undominated with respect to* $M_{-i}^{(\lambda-1)} = \prod_{j\neq i} M_j^{(\lambda-1)}$ in the sense that there exists no $m_i' \in M_i$ such that for every $m_{-i} \in M_{-i}^{(\lambda-1)}$,

$$(x(m_i', m_{-i}), t_i(m_i', m_{-i}), (m_i', m_{-i})) \succ_i (x(m), t_i(m), m).^2$$

Let $M^{(\lambda)} = \prod_{i\in\{1,2,3\}} M_i^{(\lambda)}$ and $M^{(\infty)} = \bigcap_{\lambda=0}^{\infty} M^{(\lambda)}$. A message profile $m \in M$ is said to be *iteratively undominated* $(G,\underset{\sim}{\succ})$ if $m \in M^{(\infty)}$. Let $m_i^* \in M_i$ denote the *honest message* for agent i where $m_{i,k}^* = a^*$ for all $k \in \{1,...,K\}$. The honest message profile $m^* = (m_i^*)_{i\in\{1,2,3\}} \in M$ induces the desirable alternative a^* with no monetary transfers, i.e.,

$$x(m^*)[a^*] = 1 \text{ and } t_i(m^*) = 0 \text{ for all } i \in \{1,2,3\}.$$

Theorem 1: *The honest message profile $m^* \in M$ is uniquely iteratively undominated in $(G,\underset{\sim}{\succ})$ if Conditions 1 and 2 hold.*

[2] We eliminate only *strictly* dominated messages by using the same method that was used in the studies for *virtual* implementation by Abreu and Matsushima (1992a, 1992b). Abreu and Matsushima (1994) investigated *exact* implementation, just like this paper does; however, unlike this paper, they used iteratively weakly undominated strategies where only *weakly* dominated strategies were eliminated.

Proof: Since $x(m)$ and $t_i(m)$ are independent of $m_{i,1}$, it follows from Condition 1 that each agent $i \in \{1,2,3\}$ has an incentive to announce $m_{i,1} = a^*$. Fix $h \in \{1,...,K-1\}$ and $m \in M$ arbitrarily and suppose that

$$m_{i,h'} = a^* \text{ for all } i \in \{1,2,3\} \text{ and all } h' \in \{1,...,h-1\}.$$

First, consider agent $i \in \{2,3\}$. Suppose that $m_{i,h} \neq a^*$. Let $m_i' \in M_i$ be the message for agent i such that $m_{i,h}' = a^*$ and $m_{i,h'}' = m_{i,h'}$ for all $h' \in \{1,...,K\}/\{h\}$. If

$$m_{j,h} = a^* \text{ for all } j \neq i,$$

then $x(m)$ is independent of $m_{i,h}$ and $t_i(m_i', m_{-i}) - t_i(m) \geq 0$ holds. This along with Conditions 1 and 2 implies that agent i has an incentive to announce m_i' instead of m_i. If

$$m_{j,h} \neq a^* \text{ for some } j \neq i,$$

then $t_i(m_i', m_{-i}) - t_i(m) = \varepsilon$ holds. This along with Condition 2 implies that agent i has an incentive to announce m_i' instead of m_i because $\max_{a \in A} \left| x(m)[a] - x(m_i', m_{-i})[a] \right| \leq \frac{1}{K-1}$ holds.

Next, let us consider agent 1. Suppose that $m_{1,h} \neq a^*$ and $m_{i,h} = a^*$ for each $i \in \{2,3\}$. Let $m_1' \in M_1$ be the message for agent 1 such that $m_{1,h}' = a^*$ and $m_{1,h'}' = m_{1,h'}$ for all $h' \in \{1,...,K\}/\{h\}$. Since $x(m)$ is independent of $m_{1,h}$ and $t_1(m_1', m_{-1}) = t_1(m) = 0$ holds, it follows from Condition 1 that agent 1 has an incentive to announce m_1' instead of m_1. Hence, we have proved that m^* is the unique iteratively undominated message profile.

Q.E.D.

Theorem 1 implies that by using a single *almighty* mechanism, the principal implements any alternative in iterative dominance fully, and exactly, provided that the three agents regard this alternative as being desirable. This mechanism is detail-free in a very strict sense in that it does not depend on the specification of the state space. This implies that this mechanism is independent of the social choice function that maps states to alternatives, whereas the previous mechanisms used in the implementation literature were well tailored to its fine details.

In order to implement a social choice function, it is necessary for its value to depend only on the agents' preferences. The previous works generally assumed that agents had no intrinsic preferences for honesty and that they were concerned *only* with their material interests. In this case, the principal has to invite all the relevant individuals and make them reveal their preferences, which is an extremely expensive exercise in practice. Moreover, even if it is possible for her to invite them, it might be impossible to implement the social choice function because it generally depends on not just the agents' material interests but also the factors that are irrelevant to their material interests, such as *fairness*. In contrast, in this paper, all that she is required to do for implementation is invite *three* individuals (and not all the relevant individuals) who have intrinsic preferences for honesty in a greater or lesser degree and make them announce the desirable alternative (and not the state).

To implement our mechanism, we do not even need to assume that the principal knows the set of alternatives beforehand. In fact, our mechanism can be described

thoroughly in the following simple document that she writes to the agents without mentioning the set of alternatives.

> "Tell me K number of times about what I should do. I will select one announcement profile from the last $K-1$ profiles. If two of you make the same recommendation, I will follow it. Otherwise, I will do nothing. I will impose a fine of ε if and only if you are either agent 2 or agent 3 and are one of the first to deviate from the first announcement made by agent 1."

Based on these observations, we can conclude that the presence of agents' intrinsic preferences for honesty functions effectively from the practical perspectives.

3. Private Information

Let $N = \{1,...,n\}$ denote the set of agents where $n \geq 2$. Let Ω_i denote the finite set of *private* signals for agent $i \in N$. Let $\Omega = \prod_{i \in N} \Omega_i$ denote the set of states, where $\omega = (\omega_i)_{i \in N} \in \Omega$ denote a *state*. Let $p : \Omega \to [0,1]$ denote a probability function over Ω, according to which the state is drawn randomly. A *social choice function* $f : \Omega \to A$ is defined as a mapping from states to alternatives.

The principal wants to achieve the desirable alternative $f(\omega) \in A$ that depends on $\omega \in \Omega$, which is not known to her. She delegates the alternative choice to these agents according to the following mechanism $G = (M, g, t)$, which is related to the Abreu-Matsushima mechanism with incomplete information (Abreu and Matsushima (1992b)). Let

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

Let $M_i = M_{i,1} \times \cdots \times M_{i,K}$, where $M_{i,k} = \Omega_i$ for all $k \in \{1,...,K\}$. For every $m \in M$, let

$$x(m)[a] = \frac{\#\{k \in \{\hat{K}+1,...,K\} \mid f((m_{i,k})_{i \in N}) = a\}}{K - \hat{K}} \text{ for all } a \in A,$$

for every $i \in N$,

$$t_i(m) = -\varepsilon \qquad \text{if there exist } k \in \{2,...,K\} \text{ such that } m_{i,k} \neq m_{i,1} \text{ and}$$

$$(m_{j,h})_{j \in N} = (m_{j,1})_{j \in N} \text{ for all } h \in \{1,...,k-1\},$$

and

$t_i(m) = 0$ if there exists no such k,

where $\hat{K}$ is a positive integer less than K. The principal requires each agent to announce K number of times the type of private signal that was observed. She randomly selects one announcement profile $(m_{j,k})_{j \in N}$ from the *last* $K - \hat{K}$ profiles and chooses the alternative $f((m_{j,k})_{j \in N}) \in A$, where $k \in \{\hat{K}+1, ..., K\}$. She imposes a fine of $\varepsilon > 0$ if and only if the agent is the first to deviate from her own first announcement. We assume that the monetary fine ε is as close to zero as possible.

We define a *utility function* for each agent $i \in N$ by $u_i : A \times \{-\varepsilon, 0\} \times M \times \Omega \to R$. We assume expected utility and additive separability—for every $i \in N$, there exist $v_i : A \times \Omega \to R$ and $c_i : M \times \Omega \to R$—such that

$$u_i(a, r_i, m, \omega) = v_i(a, \omega) + t_i - c_i(m, \omega).$$

It should be noted that v_i implies the utility function of agent i for her material interest and c_i is her cost function for lying.

Condition 3: For every $i \in N$, $(m, \omega) \in M \times \Omega$, and $m_i' \in M_i / \{m_i\}$,

$$c_i((m_i', m_{-i}), \omega) > c_i(m, \omega) \quad \text{if } m_{i,k}' \in \{\omega_i, m_{i,k}\} \text{ and } [m_{i,k}' = \omega_i] \Rightarrow [m_{i,k} = \omega_i]$$

for all $k \in \{1, ..., K\}$.

Condition 3 corresponds to Condition 1, which implies that each agent is sufficiently honest to dislike a white lie.

Condition 4: For every $i \in N$, $(m, \omega) \in M \times \Omega$, and $m_i' \in M_i$,

(3) $c_i((m_i', m_{-i}), \omega) - c_i(m, \omega) > \varepsilon$ if $m_{i,k} = \omega_i \neq m_{i,k}'$ for all $k \in \{1, ..., \hat{K}\}$ and

$m_{i,k} = m_{i,k}'$ for all $k \in \{\hat{K} + 1, ..., K\}$

and

(4) $(K - \hat{K})\varepsilon > \max_{(a,a',\omega,i)} \left| v_i(a, \omega) - v_i(a', \omega) \right|$.

The inequality (4) in Condition 4 corresponds to Condition 2. Condition 4 is essentially the same as the condition that for every $i \in N$, $(m, \omega) \in M \times \Omega$, and $m_i' \in M_i \setminus \{m_i\}$,

(5) $c_i((m_i', m_{-i}), \omega) - c_i(m, \omega) > \varepsilon$ if $m_{i,k} = \omega_i \neq m_{i,k}'$ for all $k \in \{1, ..., K\}$.

This implies that the utility difference between "always honest" and "always lying" is greater than the monetary fine ε. Assume that the utility for "almost always honest" ("almost always lying") is approximated by that for "always honest" ("always lying"). Then, with a sufficiently large K, we can choose an integer $\hat{K}$ such that $K - \hat{K}$ is sufficiently large to satisfy (4) but $\frac{\hat{K}}{K}$ is close to unity, which along with (5) implies (3). Since ε is chosen such that it is as close to zero as possible, the inequalities (3) in Condition 4 can be regarded as a very weak requirement for the presence of agents' preferences for honesty.

Let $u = (u_i)_{i \in N}$ denote a utility function profile. A combination (G, u) defines a *Bayesian game*. A *strategy for each agent* $i \in N$ is defined as a function $s_i : \Omega_i \to M_i$. We denote $s_i = (s_{i,k})_{k=1}^K$ and $s_i(\omega_i) = (s_{i,k}(\omega_i))_{k=1}^K$, where $s_{i,k} : \Omega_i \to \Omega_i$ and $s_{i,k}(\omega_i) \in \Omega_i$ denotes the $k-th$ announcement made by agent i. Let S_i denote the set of strategies for agent i. A strategy profile is denoted by $s = (s_i)_{i \in N}$. Let $S \equiv \prod_{i \in N} S_i$, $s(\omega) = (s_i(\omega_i))_{i \in N}$, and $s_{-i}(\omega_{-i}) = (s_j(\omega_j))_{j \in N/\{i\}}$. The solution concept is iterative dominance. Let $S_i^{(0)} = S_i$ and $S^{(0)} = \prod_{i \in N} S_i^{(0)}$. Recursively, for every $\lambda = 1, 2, \ldots$, let $S_i^{(\lambda)}$ denote the set of strategies $s_i \in S_i^{(\lambda-1)}$ for each agent i that are *undominated with respect to* $S_{-i}^{(\lambda-1)} = \prod_{j \in N/\{i\}} S_j^{(\lambda-1)}$; in other words, there exist no $m_i \in M_i$ and no $\omega_i \in \Omega_i$ such that for every $s_{-i} \in S_{-i}^{\lambda-1}$,

$$E[u_i(x(m_i, s_{-i}(\omega_{-i})), t_i(m_i, s_{-i}(\omega_{-i})), (m_i, s_{-i}(\omega_{-i})), \omega) \mid \omega_i]$$

$$> E[u_i(x(s(\omega)), t_i(s(\omega)), s(\omega), \omega) \mid \omega_i],$$

where $u_i(\alpha, r_i, m, \omega) = \sum_{a \in A} u_i(a, r_i, m, \omega)\alpha(a)$ and $E[\cdot \mid \omega_i]$ is the expectation operator given ω_i. Let $S^{(\lambda)} = \prod_{i \in N} S_i^{(\lambda)}$ and $S^{(\infty)} = \bigcap_{\lambda=0}^{\infty} S^{(\lambda)}$. A strategy profile $s \in S$ is said to be *iteratively undominated in* (G, u) if $s \in S^{(\infty)}$. We define the *honest strategy* $s_i^* \in S_i$ for agent i by

$$s_{i,k}^*(\omega_i) = \omega_i \text{ for all } k \in \{1, \ldots, K\} \text{ and all } \omega_i \in \Omega_i.$$

The honest strategy profile $s^* = (s_i^*)_{i \in N} \in S$ induces the value of the social choice function $f(\omega)$ for every state $\omega \in \Omega$ with no monetary transfers; in other words, for every $\omega \in \Omega$,

$$x(s^*(\omega))[f(\omega)] = 1 \text{ and } t_i(s^*(\omega)) = 0 \text{ for all } i \in N .$$

Theorem 2: *The honest strategy profile* $s^* \in S$ *is uniquely iteratively undominated in* (G,u) *if incentive compatibility holds; in other words, for every* $i \in N$, $\omega_i \in \Omega_i$, *and* $\omega_i' \in \Omega_i / \{\omega_i\}$,

$$E[v_i(f(\omega),\omega) \mid \omega_i] \geq E[v_i(f(\omega_i',\omega_{-i}),\omega) \mid \omega_i]. \tag{6}$$

Proof: Fix $s \in S$ and $i \in N$ arbitrarily. Fix $\omega \in \Omega$ arbitrarily. Suppose that $s_{j,k}(\omega_j) \neq s_{j,k-1}(\omega_j)$ for some $j \neq i$ and some $k \in \{2,...,\hat{K}\}$. Then, agent i is never fined at the time of announcing $m_{i,k} = \omega_i$ for all $k \in \{1,...,\hat{K}\}$. Next, suppose that $s_{j,k}(\omega_j) = s_{j,k-1}(\omega_j)$ for all $k \in \{2,...,\hat{K}\}$ and all $j \neq i$. If $s_{i,k}(\omega_i) \neq \omega_i$ for all $k \in \{1,...,\hat{K}\}$, then, by announcing $m_{i,k} = \omega_i$ for all $k \in \{1,...,\hat{K}\}$ instead, agent i can save the disutility for lying $c_i((m_i, s_{-i}(\omega_{-i})),\omega) - c_i(s(\omega),\omega)$, which is greater than ε due to (3). If $s_{i,k}(\omega_i) \neq s_{i,k-1}(\omega_i)$ for some $k \in \{2,...,\hat{K}\}$, then agent i is fined an amount ε. Since the first $\hat{K}$ announcements made by agent i never influence the alternative choice, it follows from Conditions 3 and 4 and the above arguments that agent i is willing to replace the first $\hat{K}$ announcements $(s_{i,k}(\omega_i))_{k=1}^{\hat{K}}$ with $(s_{i,k}^*(\omega_i))_{k=1}^{\hat{K}}$.

Fix $\bar{k} \in \{\hat{K}+1,...,K\}$ arbitrarily. Suppose that $s_{j,k} = s_{j,k}^*$ for all $j \in N$ and all $k \in \{1,...,\bar{k}-1\}$. Fix $\omega_i \in \Omega_i$ arbitrarily. Suppose that $s_{i,\bar{k}}(\omega_i) \neq \omega_i$. Let $m_i \in M_i$ denote the

message for agent i such that $m_{i,k} = \omega_i$ for all $k \in \{1,...,\bar{k}\}$ and $m_{i,k} = s_{i,k}(\omega_i)$ for all $k \in \{\bar{k}+1,...,K\}$.

First, suppose that $s_{j,\bar{k}}(\omega_j) \neq \omega_j$ for some $j \neq i$. Then, $t_i(s(\omega)) = -\varepsilon$ and $t_i(m_i, s_{-i}(\omega_{-i})) = 0$, which along with (4) imply that agent i prefers m_i to $s_i(\omega_i)$. Next, suppose that $s_{j,\bar{k}}(\omega_j) = \omega_j$ for all $j \neq i$. Then, $t_i(s(\omega)) = -\varepsilon$ and $t_i(m_i, s_{-i}(\omega_{-i})) \geq -\varepsilon$, which along with Condition 3 and (6) imply that agent i strictly prefers m_i to $s_i(\omega_i)$. Hence, we have proved that s^* is the unique iteratively undominated strategy profiles.

Q.E.D.

Theorem 2 implies that with minor restrictions on agents' intrinsic preferences for honesty, the principal can fully, and exactly, implement any incentive compatible social choice function in iterative dominance by using only small fines. In contrast to the previous works, we do not need any conditions, such as Bayesian monotonicity (Jackson (1991)), no consistent deception (Matsushima (1993)), and measurability (Abreu and Matsushima (1992b)), in addition to incentive compatibility. The designed mechanism is detail-free in the sense that it does not depend on the further details of the probability function and agents' utility functions for their material interests.

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.

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

Baiman, S. and B. Lewis (1989): "An Experiment Testing the Behavioral Equivalence of Strategically Equivalent Employment Contracts," *Journal of Accounting Research* 27, 1–20.

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

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

Gneezy, U. (2005): "Deception: The Role of Consequences," *American Economic Review* 95, 384–394.

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

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

Matsushima, H. (1993): "Bayesian Monotonicity with Side Payments," *Journal of Economic Theory* 45, 128–144.

Moore, J. (1992): “Implementation in Environments with Complete Information,” in *Advances in Economic Theory: Sixth World Congress*, ed. by J.J. Laffont. Cambridge University Press.

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