# Role of Honesty in Full Implementation [+]

Hitoshi Matsushima[*]

Faculty of Economics, University of Tokyo

June 4, 2007

(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, 18330035) 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, as well as a grant from the Center for Advanced Research in Finance (CARF) at the University of Tokyo. I am grateful to the anonymous referee and the associate editor for their helpful comments. All remaining errors are mine.

[*] Faculty of Economics, University of Tokyo, Hongo, Bunkyo-ku, Tokyo 113-0033, Japan

## Abstract

This paper introduces a new concept of full implementation that takes into account agents' preferences for understanding how the process works. We assume that the agents have intrinsic preferences 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 central planner. We show that the presence of such preferences functions effectively in eliminating unwanted equilibria from the practical perspective, even if the degree of the preference for honesty is small. The mechanisms designed are detail-free and involve only small fines.



**Keywords:** Intrinsic Preferences for Honesty, Detail-free Mechanisms, Full Implementation, Small Fines, Permissive Result

**JEL Classification Numbers:** C72, D71, D78, H41


# 1. Introduction

This paper introduces a new concept of full implementation that takes into account agents' preferences for understanding not just the consequence but also how the process works. We investigate Bayesian environments wherein a central planner is unaware of the desired alternative to be chosen, even if there exist multiple agents and they do receive their private signal concerning this alternative. The central planner delegates the alternative

choice to these agents by designing a mechanism, according to which each agent makes announcements about their private signals. Full implementation requires that the values of the social choice function, i.e., the desired alternatives, are induced by the *unique* Bayesian Nash equilibrium.

The previous works have constructed complicated mechanisms, which are tailored to the finer detail of specifications. This complexity makes it difficult to put the implementation theory into practice.[1] For instance, let us consider the mechanisms provided by Abreu and Matsushima (1992a, 1992b, 1994), in which the agents are required to make *multiple* announcements about their private signals at the same time. The central planner regards their first announcements as a *reference* and fines a *small* monetary amount to any agent who is the first to deviate from this reference. As long as the agents are honest during their first announcements, this device of small fines functions to incentivize the agents to keep their all other announcements honest. Incentivizing the agents to keep their first announcements honest at the outset might be a more problematic issue. In order to solve this issue, Abreu and Matsushima incorporated an additional incentive scheme into the device of small fines, which is, however, not detail-free, i.e., it depends heavily on the finer detail of specifications such as the probability function and the utility functions. The failure to make the mechanisms detail-free is the main drawback of the implementation theory from the practical perspective.[2]

---

[1] See survey articles such as Moore (1992), Palfrey (1992), Osborne and Rubinstein (1994, Chapter 10), and Maskin and Sjöström (2002).

[2] Another practical problem concerns whether actual agents play iteratively undominated strategies. In experimental economics, it is a well known idea that even if subjects never enforce many iterations of removal at once, they may learn to achieve a sufficiently large number of iterations in the long run. See

The purpose of this paper is to demonstrate the possibility of implementing social choice functions without harming the detail-free concept of mechanism design. The crucial assumption is that each of these agents has an *intrinsic preference for honesty* in the sense that she/he dislikes the idea of telling white lies that do not influence her welfare but instead go against the intention of the central planner. With the intrinsic preference for honesty in this sense, we do not need to incorporate any additional incentive scheme with the device of small fines. All we have to do is just to *keep the agents' first few announcements irrelevant to the alternative decision*. Hence, by using only detail-free mechanisms, we can fully and exactly implement any incentive compatible social choice function in iterative dominance. Apart from incentive compatibility, we do not require any condition on social choice functions. 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.[3]

Several experimental economics researches such as Gneezy (2005) emphasized that the role of intrinsic preferences in this manner is non-negligible in economic decisions. Charness and Dufwenberg (2006), on other hand, raised the alarm that agents' intrinsic preferences to influence their decisions are heavily dependent on contexts and framings. For instance, agents' intrinsic costs of lying may not be significant as long as they expect that the central planner believes that they lie. Moreover, since each agent makes so many

---

Camerer (2003), for instance. There is a difficulty in applying this idea to our situation in that each agent's own experiences are severely limited, and therefore, she/he has to utilize the other agents' experiences. Huck, Jehiel, and Rutter (2006) obtained experimental results stating that learning is affected by the framing of feedback information about the other agents' experiences. How to fix this framing in the first place is an interesting question but is beyond the purpose of this paper.

announcements at once, it is inevitable that her/his intrinsic cost of lying for each single announcement is severely limited.

Despite the fragility of the intrinsic preferences in this manner, the result of this paper should be regarded as being quite permissive. Overcoming this fragility only requires that *the proportion of announcements that are irrelevant to the alternative decision be sufficient.* This would show that even if each agent's intrinsic cost of lying for *all* of her/his announcements is close to zero, any incentive compatible social choice function is fully and exactly implementable in iterative dominance.[4]

This paper is organized as follows. Section 2 defines the model. Section 3 specifies detail-free mechanisms. Section 4 shows the main theorem.

## 2. The Model

Let $A$ denote a finite set of alternatives; $\Delta$, the set of lotteries over the alternatives; and $N = \{1,...,n\}$, a finite set of agents, where $n \geq 2$. Further, let $\Omega_i$ denote a finite set of *private* signals for agent $i \in N$, where we set $\omega_i \in \Omega_i$; $\Omega = \prod_{i \in N} \Omega_i$, the set of private signal profiles; and $p : \Omega \to [0,1]$, a probability function over $\Omega$, according to which the private

[3] There are exceptions such as Glazer and Rubinstein (1998) and Eliaz (2002).
[4] We eliminate only *strictly* dominated messages by using the same method 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.

signal profile $\omega = (\omega_i)_{i \in N} \in \Omega$ is drawn randomly. A *social choice function* $f : \Omega \to A$ is defined as a mapping from private signal profiles to alternatives.

The central planner wants to achieve the desirable alternative $f(\omega) \in A$ that depends on the private signal profile $\omega \in \Omega$, which is not known to her/him. She/he delegates the alternative choice to the agents according to a mechanism $G = (M, g, t)$, where $M = \prod_{i \in N} M_i$, $m_i \in M_i$, $M_i$ is a finite set of messages for each agent $i$, $x : M \to \Delta$, $t = (t_i)_{i \in N}$, and $t_i : M \to R$. When the agents announce a message profile $m = (m_i)_{i \in N} \in M$, the central planner 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. We focus on mechanisms, in which each agent makes *multiple* announcements about her/his private signal; a positive integer $K$ exists such that for every $i \in N$,

$$M_i = \Omega_i^K \text{ and } M_i = M_{i,1} \times \cdots \times M_{i,K},$$

where $m_i = (m_{i,k})_{k=1}^K \in M_i$, $M_{i,k} = \Omega_i$, and $m_{i,k} \in M_{i,k}$ for all $k \in \{1, ..., K\}$. For every $k \in \{1, ..., K\}$, we term $m_{i,k} \in M_{i,k}$ as the $k-th$ announcement of agent $i$.

We define a *utility function* for each agent $i \in N$ by $u_i : A \times R \times M \times \Omega \to R$, where there exist functions $v_i : A \times \Omega \to R$ and $c_i : [0,1] \times \Omega_i \to R$ such that

$$c_i(0, \omega_i) = 0,$$

$$c_i(r, \omega_i) \text{ is continuous and increasing with respect to } r \in [0,1],$$

and

$$u_i(a,t_i,m,\omega) = v_i(a,\omega) + t_i - c_i(\frac{\#\{k \in \{1,...,K\} \mid m_{i,k} \neq \omega_i\}}{K}, \omega_i).$$

Note that $c_i(\frac{\#\{k \in \{1,...,K\} \mid m_{i,k} \neq \omega_i\}}{K}, \omega_i)$ implies agent $i's$ intrinsic cost of lying when she/he receives the private signal $\omega_i \in \Omega_i$ and announces the message $m_i \in M_i$. This intrinsic cost depends on the *proportion* of her/his dishonest announcements, $\frac{\#\{k \in \{1,...,K\} \mid m_{i,k} \neq \omega_i\}}{K}$. This cost does *not* depend on the absolute number of dishonest announcements. Hence, the intrinsic cost of lying for each single announcement is severely limited whenever the number of announcements that each agent is required to make is very large. Further, note that $v_i(a,\omega)$ implies the utility of agent $i$ for her/his material interest. Moreover, we assume quasi-linearity and risk-neutrality in terms of monetary transfers.

We shall confine our attention to social choice functions $f$ that satisfy *incentive compatibility* in terms of the agents' material interests in that for every $i \in N$, $\omega_i \in \Omega_i$, and $\omega_i' \in \Omega_i / \{\omega_i\}$,

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

where $E[\cdot \mid \omega_i]$ is the expectation operator given $\omega_i$. Incentive compatibility implies that truth-telling is a Bayesian Nash equilibrium in the direct mechanism irrespective of whether or not the agents have intrinsic 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 of 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 used in this paper is iterative dominance, which is defined as follows. Let $S_i^{(0)}=S_i$ and $S^{(0)}=\prod_{i\in N}S_i^{(0)}$. Recursively, for every $h=1,2,\ldots$, let $S_i^{(h)}$ denote the set of strategies $s_i\in S_i^{(h-1)}$ for each agent $i$ that are *undominated with respect to* $S_{-i}^{(h-1)}=\prod_{j\in N/\{i\}}S_j^{(h-1)}$; in other words, there exist no $m_i\in M_i$ and no $\omega_i\in\Omega_i$ such that for every $s_{-i}\in S_{-i}^{(h-1)}$,

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

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

where we denote $u_i(\alpha,r_i,m,\omega)=\sum_{a\in A}u_i(a,r_i,m,\omega)\alpha(a)$ for each $\alpha\in\Delta$. Let $S^{(h)}=\prod_{i\in N}S_i^{(h)}$ and $S^{(\infty)}=\bigcap_{h=0}^{\infty}S^{(h)}$. 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 $\omega\in\Omega$ with no monetary transfers; in other words, for every $\omega\in\Omega$,

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

## 3. Mechanism Design

We fix a positive real number $\varepsilon > 0$ such that

(2) $\varepsilon < c_i(1, \omega_i)$ for all $i \in N$,

which implies that $\varepsilon$ is selected less than the intrinsic disutility for each agent when she/he lies during *all* her/his announcements. Note that there exists such an $\varepsilon$, because $c_i(r_i, \omega_i)$ is continuous and increasing with respect to $r_i \in [0,1]$, and $c_i(0, \omega_i) = 0$. Moreover, we fix two positive integers $K$ and $\hat{K}$ such that $K > \hat{K}$,

(3) $\varepsilon < c_i(r + \frac{\hat{K}}{K}, \omega_i) - c_i(r, \omega_i)$ for all $r \in [0, 1 - \frac{\hat{K}}{K}]$,

and

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

From inequality (2) and the continuity and increase of $c_i$, note that there exist such a $K$ and $\hat{K}$. In fact, by fixing $K$ as sufficiently large, we can choose $\hat{K}$ to satisfy the following two properties:

$\frac{\hat{K}}{K}$ is sufficiently close to unity to satisfy inequality (3),

and

$K - \hat{K}$ is sufficiently large to satisfy inequality (4).

Based on $(\varepsilon, K, \hat{K})$ defined above, we specify the mechanism, denoted by $G(\varepsilon, K, \hat{K}) = (M, g, t)$ as follows; for every $m \in M$,

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

for every $i \in N$,

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

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

and

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

The central planner requires each agent to announce $K$ number of times the type of private signal that was observed. She/he randomly selects one announcement profile $(m_{j,k})_{j\in N} \in \underset{i\in N}{\times} M_{j,k}$ 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/he imposes a fine of $\varepsilon > 0$ if and only if the agent is the first to deviate from her own first announcement.

Note that the early $\hat{K}$ announcement profiles, i.e., $(m_{j,k})_{j\in N}$ for $k \in \{1,...,\hat{K}\}$, are irrelevant to the alternative decision $x(m)$.[5] The mechanism $G(\varepsilon, K, \hat{K})$ involves only small fines given by $\varepsilon > 0$.

Note that the mechanism $G(\varepsilon, K, \hat{K})$ is detail-free in the following sense. Let us select $\varepsilon > 0$ as close to zero, a positive real number $\lambda \in (0,1)$ as being close to unity, and a positive real number $Q > 0$ such that it is sufficiently large. Moreover, let us select $K$ and

[5] Note that these profiles are relevant to the monetary transfers $(t_i(m))_{i\in N}$.

$\hat{K}$ to be sufficiently large, such that $\frac{\hat{K}}{K}$ is greater than $\lambda$ and $(K-\hat{K})\varepsilon$ is greater than $Q$.

Note that inequalities (3) and (4) hold whenever

$$c_i(r+\lambda,\omega_i)-c_i(r,\omega_i)\geq\varepsilon \text{ for all } r\in[0,1-\lambda],$$

and

$$\max_{(a,a',\omega,i)\in A^2\times\Omega\times N}\left|v_i(a,\omega)-v_i(a',\omega)\right|\leq Q,$$

which are very weak restrictions because $\varepsilon$ is selected such that it is close to zero and $Q$ is selected such that it is sufficiently large. Hence, we can say that $G(\varepsilon,K,\hat{K})$ does not depend on the finer details of specifications such as the probability function and the utility functions.[6]

## 4. Main Theorem

The following theorem shows that with incentive compatibility, truth-telling is the unique iteratively undominated strategy profile in $(G(\varepsilon,K,\hat{K}),u)$, which implies that any incentive compatible social choice function is fully implementable in iterative dominance. 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.

[6] The construction of $G(\varepsilon,K,\hat{K})$ depends on the social choice function $f$. Needless to say, $G(\varepsilon,K,\hat{K})$ functions rely crucially on the incentive compatibility of the social choice function.

Inequality (3) guarantees that each agent is willing to keep her/his early $\hat{K}$ announcements honest because her/his intrinsic cost of lying for all of these announcements is greater than the small monetary fine $\varepsilon$. Given that the early $\hat{K}$ announcements are honest, inequality (4), along with incentive compatibility, guarantees that the device of small fines functions in incentivizing each agent to keep her/his latter $K-\hat{K}$ announcements honest in the same manner as in Abreu and Matsushima (1992a, 1992b, 1994).

**The Theorem:** *The honest strategy profile* $s^* \in S$ *is uniquely iteratively undominated in* $(G(\varepsilon, K, \hat{K}), u)$.

**Proof:** Fix $s \in S$ and $i \in N$ arbitrarily. Further, fix $\omega \in \Omega$ arbitrarily. Suppose that

$$s_{j,k}(\omega_j) \neq s_{j,k-1}(\omega_j) \text{ for some } j \neq i \text{ and some } k \in \{2,...,\hat{K}\}.$$

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

$$s_{j,k}(\omega_j) = s_{j,k-1}(\omega_j) \text{ for all } k \in \{2,...,\hat{K}\} \text{ and all } j \neq i.$$

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

$$c_i\left(\frac{\#\{k \in \{1,...,K\} \mid s_{i,k}(\omega_i) \neq \omega_i\}}{K}, \omega_i\right) - c_i\left(\frac{\#\{k \in \{1,...,K\} \mid s_{i,k}(\omega_i) \neq \omega_i\} - \hat{K}}{K}, \omega_i\right),$$

which is greater than $\varepsilon$ 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 $\varepsilon$. Since the early $\hat{K}$ announcements of agent $i$ do not influence the alternative decision, it follows that agent $i$ is willing to replace the early $\hat{K}$ 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

$$s_{j,k} = s^*_{j,k} \text{ for all } j \in N \text{ and all } k \in \{1,...,\bar{k}-1\}.$$

Further, 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

$$s_{j,\bar{k}}(\omega_j) \neq \omega_j \text{ for some } j \neq i.$$

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

$$s_{j,\bar{k}}(\omega_j) = \omega_j \text{ for all } j \neq i.$$

Then, $t_i(s(\omega)) = -\varepsilon$ and $t_i(m_i, s_{-i}(\omega_{-i})) \geq -\varepsilon$, which, along with the intrinsic preferences for honesty and incentive compatibility given by inequality (1), 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.**

# 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.

Camerer, C. (2006): *Behavioral Game Theory*, Russell Sage Foundation.

Charness, G. and M. Dufwenberg (2006): "Promises and Partnerships," *Econometrica* 74, 1579–1601.

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.

Huck, S., P. Jehiel, and T. Rutter (2006): "Information Processing, Learning, and Analogy-Based Expectations: An Experiment," mimeo.

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.