Behavioral Aspects of Implementation Theory[+]

Hitoshi Matsushima[*]
Faculty of Economics, University of Tokyo

September 26, 2007

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

Abstract

This paper incorporates behavioral economics into implementation theory. We use mechanisms that are strictly detail-free. We assume that each agent dislikes telling a white lie when such lying does not serve her/his material interest. We present a permissive result wherein by using just a single detail-free mechanism, any alternative can be uniquely implemented in iterative dominance as long as the agents regard this alternative as being socially desirable.

Keywords: Behavioral Economics, Implementation Theory, White Lie Aversion, Detail-Freeness, Possibility Theorem.
JEL Classification Numbers: C72, D71, D78, H41

1. Introduction

This paper investigates the implementation problem in which the central planner wishes to choose the socially desirable alternative, although she/he is not aware of this alternative. We assume that there are three agents who are correctly aware of this alternative. The central planner delegates the alternative choice to these agents by requesting them to make honest announcements. The central issue of the implementation theory is whether their honest announcements can be supported by the unique Nash equilibrium in a decentralized procedure. For previous works in the implementation literature, see the surveys conducted by Moore (1992), Osborne and Rubinstein (1994), and Maskin and Sjöström (2002).

From the viewpoint of behavioral economics, any agent to whom the central planner delegates the alternative choice should be expected to take her/his social preference into account. For instance, she/he may dislike failing to measure up to the expectations of the central planner because of her/his *guilt aversion*. This tendency is supported by the laboratory experiments conducted by Gneezy (2005) and Charness and Dufwenberg (2006). However, barring a few works such as Glazer and Rubinstein (1998) and Eliaz (2002), previous works did not consider this behavioral aspect. For details on behavioral game theory in general, see the survey conducted by Camerer (2006).

In order to incentivize the agents, the central planner has to carefully design a mechanism. From the practical viewpoint, the designed mechanism should be *free from the details* of the model specifications. However, previous works designed mechanisms that were heavily dependent on the state space, the social choice function that maps states to alternatives, and the state-dependent utility functions. If we overlook the agents' social preferences and focus only on their material interests, it will be impossible to implement the socially desirable alternatives by merely designing detail-free mechanisms.

This paper demonstrates that incorporating behavioral economics into the implementation theory is an effective method from the practical viewpoint. Let us consider

a mechanism à la Abreu and Matsushima (1992, 1994), which is described in the following message that the central planner conveys to the three agents, i.e., agents 1, 2, and 3.

> I request each agent to tell me a sufficiently large K number of times about what I should do for the social benefit. I will randomly select one announcement profile from these K profiles. If two or three agents make the same recommendation, I will follow it; otherwise I will do nothing**.** I also request agent 1 to tell me once more about what I should do. By using the $(K+1)-th$ announcement as reference, I wish to identify and penalize liars. I will impose a small fine of $\varepsilon > 0$ if and only if you are agent 2 or agent 3 and are one of the last agents to deviate from this $(K+1)-th$ announcement.

This mechanism is entirely free from the specifications of the state space, the social choice function, the utility functions, and even the set of alternatives. Further, this mechanism has severe multiplicity of the Nash equilibria as long as the agents are motivated merely by their material interests. With respect to alternatives, the constant announcement of an alternative by all the agents results in the Nash equilibrium in this case.

This paper argues that this multiplicity is not robust to the agent's slight behavioral motives. As long as the other agents follow this Nash equilibrium, agent 1's announcements do not influence the alternative choice and the monetary transfer to her/him, that is, her/his material interests. Further, her/his announcements merely influence the other agents' material interests, because only small fines are permitted. In this case, agent 1 should be expected not to speak any *white lie* because of her/his guilt aversion named "*white lie aversion.*" This along with iterative removals is the driving force behind eliminating dishonest announcements.

Each agent should display some inclination to engage in white lie aversion; such inclination is all that is needed for this mechanism to function. The possibility theorem is very permissive—*given the presence of a slight white lie aversion, any alternative is uniquely implemented in iterative dominance as long as the agents regard this alternative as being socially desirable.*

The present paper does not require any restriction on their preferences, with the exception of a naïve form of white lie aversion. In contrast, Matsushima (2007) investigated incomplete information by assuming quasi-linearity and expected utility. The mechanisms designed in Matsushima (2007) depended on the social choice function.

This paper is organized as follows. Section 2 presents the model. Section 3 defines iterative dominance, and Section 4 defines white lie aversion and depicts the main theorem.

2. The Model

We consider a situation wherein the central planner is not aware of the socially desirable alternative, whereas agents 1, 2, and 3 are aware of it. The central planner delegates the alternative choice to these agents using the following procedure. The central planner requests each agent to announce the alternative K number of times, following which she/he randomly selects one announcement profile from among these K profiles. Here, $K > 0$ is a sufficiently large positive integer. If at least two agents announce the same alternative, then she/he chooses that alternative. In the absence of such an alternative, she/he chooses the status quo given by $\bar{a} \in A$, where A denotes the set of alternatives. Further, the central planner requests agent 1 to announce the alternative once more as the $(K+1)-th$ announcement. The central planner imposes a fine $\varepsilon > 0$ if and only if the agent is agent 2 or agent 3 and is one of the last agents to deviate from this $(K+1)-th$ announcement. We permit only small fines, i.e., ε is close to zero. The $(K+1)-th$ announcement of agent 1 does not influence the alternative choice. Agent 1 is never fined.

The above procedure is described by the following mechanism: $G = (M, x, t)$. Let M_i denote the set of *messages* for each agent i. Specify

$$M_1 = A^{K+1},\ M_2 = A^K, \text{ and } M_3 = A^K.$$

Let $M_1 = \underset{k=1}{\overset{K+1}{\times}} M_{1,k}$, $M_2 = \underset{k=1}{\overset{K}{\times}} M_{2,k}$, and $M_3 = \underset{k=1}{\overset{K}{\times}} M_{3,k}$, where $M_{i,k} = A$. Let $m_i = (m_{i,k})$, where $m_{i,k} \in M_{i,k}$ denotes the $k-th$ announcement of agent i. Let $M = M_1 \times M_2 \times M_3$ and $m = (m_1, m_2, m_3) \in M$. For every $k \in \{1, ..., K\}$, let $(m_{1,k}, m_{2,k}, m_{3,k}) \in A^3$ denote the $k-th$ announcement profile.

A *simple lottery* over alternatives is defined as $\alpha : A \to [0,1]$, which has a countable subset $\Gamma \subset A$ such that $\alpha(a) > 0$ for all $a \in \Gamma$, $\alpha(a) = 0$ for all $a \notin \Gamma$, and $\sum_{a \in \Gamma} \alpha(a) = 1$. Let Δ denote the set of simple lotteries. Let $x : M \to \Delta$, $t = (t_i)_{i \in N}$, and $t_i : M \to \{-\varepsilon, 0\}$ for all $i \in \{1, 2, 3\}$. When the agents announce a message profile $m \in M$, the central planner

chooses any alternative $a \in A$ with probability $x(m)(a) \in [0,1]$ and certainly makes a monetary transfer $t_i(m) \in \{-\varepsilon, 0\}$ to each agent i.

For every $m \in M$, specify

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

and

$$x(m)(\bar{a}) = 1 - \sum_{a \neq \bar{a}} x(m)(a).$$

For every $k \in \{1,...,K\}$, the central planner selects the $k-th$ announcement profile $(m_{1,k}, m_{2,k}, m_{3,k}) \in A^3$ with probability $\frac{1}{K}$ and chooses any alternative $a \in A$ when at least two agents announce this alternative, i.e.,

$$m_{i,k} = a \text{ for at least two agents } i \in \{1,2,3\}.$$

In the absence of such an alternative, she/he chooses the status quo $\bar{a}$.

For every $m \in M$, specify

$$t_1(m) = 0,$$

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

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

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

and

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

Each agent $i \in \{2,3\}$ is fined if and only if she/he is one of the last agents to deviate from the $(K+1)-th$ announcement $m_{1,K+1} \in A$ of agent 1. Agent 1 is never fined.

This mechanism does not depend on the specifications of the state space, the social choice function, and the utility functions.

3. Iterative Dominance

A preference for each agent $i \in \{1,2,3\}$ is defined as an ordering $\underset{\sim i}{\succ}$ on $\Delta \times \{-\varepsilon, 0\} \times M$. Here, $(\alpha, r_i, m) \underset{\sim i}{\succ} (\alpha', r_i', m')$ implies that agent i does not prefer (α', r_i', m') to (α, r_i, m), and $(\alpha, r_i, m) \succ_i (\alpha', r_i', m')$ implies that agent i strictly prefers (α, r_i, m) to (α', r_i', m'), that is, $(\alpha, r_i, m) \underset{\sim i}{\succ} (\alpha', r_i', m')$ and $\sim[(\alpha', r_i', m') \underset{\sim i}{\succ} (\alpha, r_i, m)]$. Further, $(\alpha, r_i, m) \sim_i (\alpha', r_i', m')$ implies that agent i is indifferent between (α, r_i, m) and (α', r_i', m'), that is, $(\alpha, r_i, m) \underset{\sim i}{\succ} (\alpha', r_i', m')$ and $(\alpha', r_i', m') \underset{\sim i}{\succ} (\alpha, r_i, m)$.

Since the negative transfer $-\varepsilon$ harms an agent's welfare, it is appropriate to assume that for every $i \in \{1,2,3\}$ and $(\alpha, m) \in \Delta \times M$,

$$(\alpha, 0, m) \succ_i (\alpha, -\varepsilon, m).$$

Given that K is sufficiently large, it is appropriate to assume that for every $i \in \{1,2,3\}$, $(\alpha, m) \in \Delta \times M$, and $\alpha' \in \Delta$,

(1) $$(\alpha, 0, m) \succ_i (\alpha', -\varepsilon, m), \text{ if } \sum_{a \in \Gamma \cup \Gamma'} \left|\alpha(a) - \alpha'(a)\right| \leq \frac{1}{K},$$

where Γ and Γ' denote the supports of the simple lotteries α and α', respectively. Since K is sufficiently large, assumption (1) implies that as long as the difference between the simple lotteries is small enough, each agent prefers no fine to the negative transfer $-\varepsilon$.

Let $\underset{\sim}{\succ} \equiv (\underset{\sim 1}{\succ}, \underset{\sim 2}{\succ}, \underset{\sim 3}{\succ})$ denote a preference profile. 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 $\lambda = 1, 2, \ldots$, let $M_i^{(\lambda)}$ denote the set of messages $m_i \in M_i^{(\lambda-1)}$ for each agent i, which 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 $m_{-i} \in M_{-i}^{(\lambda-1)}$,

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

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 *uniquely iteratively undominated* in $(G, \succ)$ if $M^{(\infty)} = \{m\}$.

Arbitrarily fix any alternative $a^* \in A$, which is regarded as the *socially desirable* alternative. Let $m_i^* = (m_{i,k}^*) \in M_i$ denote the *honest message* for agent i, where $m_{i,k}^* = a^*$ for all k. The honest message profile $m^* = (m_1^*, m_2^*, m_3^*) \in M$ induces the socially desirable alternative a^* with no monetary transfers, i.e.,

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

4. White Lie Aversion

We introduce a condition on $\succsim$ named *white lie aversion* that requires each agent to dislike telling a lie as long as this lie does not influence the alternative choice and the monetary transfer to her/him.

White Lie Aversion: For every $(\alpha, r_1, m) \in \Delta \times \{-\varepsilon, 0\} \times M$ and every $m_1' \in M_1 \setminus \{m_1\}$,

(2) $(\alpha, r_1, (m_1', m_{-1})) \succ_1 (\alpha, r_1, m)$ if $m_{1,k}' \in \{a^*, m_{1,k}\}$ and $[m_{1,k} = a^*] \Rightarrow [m_{1,k}' = a^*]$ for all $k \in \{1, ..., K+1\}$.

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

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

Note that agent 1 strictly prefers not to speak white lies, while this is not necessarily the case with agents 2 and 3.

The Theorem: *Under white lie aversion, the honest message profile $m^* \in M$ is uniquely iteratively undominated in $(G, \succsim)$.*

Proof: Since $x(m)$ and $t_1(m)$ are independent of $m_{1,K+1}$, it follows from (2) that agent 1 has a strict incentive to announce $m_{i,K+1} = a^*$. Arbitrarily fix $k \in \{1, ..., K\}$ and $m \in M$ and suppose that

$$m_{i,K+1} = a^*$$

and

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

Let us consider agent $i \in \{2,3\}$. Suppose $m_{i,k} \neq a^*$. Let $m_i' \in M_i$ be the message for agent i such that

$$m_{i,k}' = a^* \text{ and } m_{i,h}' = m_{i,h} \text{ for all } h \neq k.$$

If

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

then $x(m)$ is independent of $m_{i,k}$ and $t_i(m_i', m_{-i}) - t_i(m) \geq 0$ holds. This along with (1) and (3) implies that agent i has a strict incentive to announce m_i' instead of m_i. If

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

then $t_i(m_i', m_{-i}) - t_i(m) = \varepsilon$ holds. This along with (1) and (3) implies that agent i has a strict incentive to announce m_i' instead of m_i, because $\sum_{a \in \Gamma \cup \Gamma'} \left| x(m)(a) - x(m_i', m_{-i})(a) \right| \leq \frac{1}{K}$ holds, where Γ and Γ' denote the supports of $x(m)$ and $x(m_i', m_{-i})$, respectively.

Let us consider agent 1. Suppose

$$m_{1,k} \neq a^* \text{ and } m_{i,k} = a^* \text{ for each } i \in \{2,3\}.$$

Let $m_1' \in M_1$ be the message for agent 1 such that

$$m_{1,k}' = a^* \text{ and } m_{1,h}' = m_{1,h} \text{ for all } h \neq k.$$

Since $x(m)$ is independent of $m_{1,k}$ and $t_1(m_1', m_{-1}) = t_1(m) = 0$ holds, it follows from (2) that agent 1 has a strict incentive to announce m_1' instead of m_1. Hence, we have proved that m^* is uniquely iteratively undominated. **Q.E.D.**

References

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

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.

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 (2007): "Role of Honesty in Full Implementation," forthcoming in *Journal of Economic Theory*.

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.