UTMD Working Paper

UTMD-002

Implementation, Honesty, and Common Knowledge

Hitoshi Matsushima
University of Tokyo

This Version: December 22, 2020

Implementation, Honesty, and Common Knowledge[1]

Hitoshi Matsushima[2]
University of Tokyo

December 22, 2020

Abstract

We investigate the unique implementation problem of social choice functions (SCFs) from ethical and epistemological concerns. According to Matsushima and Noda (2020), we consider the possibility that in higher-order beliefs there exists an agent who is honest, that is, motivated by intrinsic preference for honesty as well as material interest. We assume only weak honesty in that an honest agent is mostly motivated by material interests and even tells a white lie. We show a very permissive result that any social choice function is uniquely implementable in Bayes Nash equilibrium if "all agents are selfish" never happens to be common knowledge. Hence, any SCF is uniquely implementable even if all agents are selfish and "all agents are selfish" is mutual knowledge. Importantly, any ethical SCF is uniquely implementable whenever "all agents are selfish" never happens to be common knowledge while it is never uniquely implementable otherwise.

Keywords: ethical social choice function, unique implementation, weak honesty, common knowledge on selfishness, permissive result.
JEL Classification Numbers: C72, D71, D78, H41

---

[1] This study was supported by a grant-in-aid for scientific research (KAKENHI 20H00070) 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. I am grateful to Professor Shunya Noda for useful comments. All errors are mine.
[2] Department of Economics, University of Tokyo, Hongo, Bunkyo-ku, Tokyo 113-0033, Japan. E-mail: hitoshi @ e.u-tokyo.ac.jp

1. Introduction

This study investigates the unique implementation problem of social choice functions (SCFs) from ethical and epistemological concerns. A central planner attempts to implement the desirable allocation implied by an SCF in a contingent manner on the state. She (or he) does not know the state, while there exist multiple agents who are fully informed of it (*complete information concerning the state*). The central planner attempts to incentivize these agents to announce about the state sincerely by designing a decentralized mechanism that consists of message spaces, an allocation rule, and a payment rule. The question is, under what condition can the central planner implement the SCF via unique equilibrium behavior?

We assume that each agent is either selfish or honest. A selfish agent is only concerned about the material utility, while an honest agent is concerned about the intrinsic preference for honesty as well. However, importantly, we do not assume any possibility that there exists an honest agent as a participant in the central planner's problem. We instead consider an epistemological possibility that an honest agent exists, not in the mechanism, but in the participants' higher-order beliefs (*incomplete information concerning selfishness and honesty*). By considering a type space concerning such beliefs, we demonstrate a very permissive result: Any SCF is uniquely implementable in Bayes Nash equilibrium if "all agents are selfish" never happens to be common knowledge.

Matsushima and Noda (2020), which is a companion paper, investigated an information elicitation problem and showed that a device of quadratic scoring rule can incentivize agents to announce truthfully as unique equilibrium behavior if and only if "all agents are selfish" never happens to be common knowledge. Matsushima and Noda however did not consider what the central planner use the information for. This study explicitly considers the central planner's purpose as a SCF and extends Matsushima and Noda (2002) to the general implementation problem.

This study is in contrast with previous literature on implementation theory, which commonly and implicitly assumed that "all agents are selfish" is common knowledge and focused on the consideration of SCFs that are *material*, that is, depend only on agents' material utilities (Arrow, 1951; Hurwicz, 1972; Gibbard, 1973; Satterthwaite,

1975; Abreu and Matsushima, 1992; Maskin, 1977/1999).[3] With this common knowledge assumption, it is impossible in principle to uniquely implement any SCF that is not material but *ethical*, that is, depends not only on an agent's material utilities but also on factors that have nothing to do with their material utilities, such as equity, fairness, and equality concerns.[4] This study suggests a highly positive potential for implementing such SCFs.

In this respect, Matsushima (2008a; 2008b) is the pioneering work, which assumed that agents are either selfish or honest and then showed that any SCF, whether material or ethical, is uniquely implementable in Nash equilibrium whenever there exist honest agents. Following Matsushima, many studies such as Dutta and Sen (2012), Kartik, Tercieux, and Holden (2014), Saporiti (2014), Ortner (2015), and Mukherjee, Muto, and Ramaekers (2017) showed their respective possibility theorems by introducing (partial) honesty into the implementation problem.[5]

By applying Matsushima and Noda (2020), we make significant progress in this new research trend in the following two points. First, the previous works assumed that there exists an honest agent as a participant in the mechanism, at least with a positive probability, while this study does not assume it at all: we permit that all agents are selfish and "all agents are selfish" is mutual knowledge. We only rule out the case in which "all agents are selfish" is common knowledge.

Second, the previous works assumed that an honest agent never tells a white lie, that is, a lie that does not influence her material utility. The methods of mechanism design in these studies depended crucially on this assumption. However, given that real people may be more or less influenced by various irrational motives, we must say that this assumption is very restrictive.

In contrast, this study permits even honest agents to tell white lies: an honest agent feels guilty about lying only when this lying increases the agent's material benefit. This

---

[3] For surveys of implementation theory, see Moore (1992), Jackson (2001), Palfrey (2002), and Maskin and Sjöström (2002).

[4] An exception is Matsushima (2019), which assumed that the state is ex-post verifiable and proved that any SCF, whether material or ethical, is uniquely implementable even if "all agents are selfish" is common knowledge. This study does not assume such verifiability.

[5] See also Matsushima (2013), Korpela (2012; 2014), Yadav (2016), Lombardi and Yoshihara (2017; 2018; 2019), Dogan (2017), and Savva (2018).

assumption is consistent with empirical and experimental studies (Abeler, Nosenzo, and Raymond, 2019).[6]

This study only requires *weak honesty*: an honest agent is almost motivated by material interest. An honest agent can tell a white lie and does not even exist in the mechanism. To make full use of such weak honesty, according to Matsushima and Noda (2020), we design a part of the payment rule for each agent as a quadratic scoring rule (Brier, 1950). The quadratic scoring rule can set aside various non-selfish motives of the agent and prioritize the agent's monetary interest to announce the same as the other. However, as Abeler, Nosenzo, and Raymond (2019) point out, the intrinsic preference for honesty remains unexcluded, and an honest agent still prefers announcing a little more honestly than selfish agents. This will be the driving force for a tail-chasing competition through which each agent announces more honestly than the other, reaching the point at which all agents report honestly. Matsushima and Noda (2020) proved in an information elicitation problem that this tail-chasing competition functions as long as "all agents are selfish" never happens to be common knowledge. By integrating this finding with a standard method of mechanism design explored by Abreu and Matsushima (1992) in a semantical manner, we prove the permissive result in the unique implementation problem. The designed mechanism is bounded, budget-balancing, and satisfies limited solvency, that is, it only uses small side payments.

This study will bring hope to central planners who lack the information necessary for normative judgments such as "are social benefits fairly distributed in the society?", "who needs relief from poverty?", and "how will decision-making affect outsiders and future generations?". Selfish people are generally unmotivated by such ethical concerns as part of their material interest, even if they have keen interests in such ethical concerns and are even knowledgeable about them. The common knowledge assumption on selfishness implies that society is divided into a group of selfish people and a group of honest people and these groups are disconnected with each other. With this assumption, the central planner cannot derive the ethical information from selfish people correctly.

[6] Abeler, Nosenzo, and Raymond (2019) provided a detailed meta-analysis: by using data from 90 studies involving more than 44,000 subjects across 47 countries, they showed that subjects who were in trade-offs between material interest and honesty gave up a large fraction of potential benefits from lying.

However, if selfish people and honest people are epistemologically connected with each other in higher-order beliefs, the central planner can properly derive such information from selfish agents and reflect it in her normative judgment, that is, uniquely implement any ethical SCF she desires.

The study is organized as follows. Section 2 presents the model. Section 3 semantically defines a class of indirect mechanisms and intrinsic preference for honesty. Section 4 defines the unique implementation in Bayes Nash equilibrium and states the main theorem. Section 5 explains the design of the mechanism for the proof of the main theorem and Section 6 outlines the logic behind this proof. Section 7 defines ethical SCFs and material SCFs and demonstrates a necessary and sufficient condition for the unique implementation of ethical SCFs. Section 8 concludes. The Appendix shows the full proof of the main theorem.

2. The Model

This study investigates a situation in which a central planner attempts to achieve a desirable allocation in a contingent manner on the state. Let $N \equiv \{1,...,n\}$ denote the finite set of all agents, where $n \geq 3$. Let A denote the non-empty and finite set of all allocations. Let Ω denote the non-empty and finite set of all states. The *social choice function* (SCF) is defined as $f : \Omega \to \Delta(A)$.[7] For every $\omega \in \Omega$, $f(\omega) \in \Delta(A)$ implies the desirable (distribution of) allocation at the state ω.[8]

We assume that the central planner does not know the state, while all agents are fully informed of it. Each agent is either *selfish* or *honest*. (More details on the meaning of selfishness and honesty will be subsequently explained.) No agent knows if the other agents are selfish or honest. To describe agents' higher-order beliefs concerning selfishness and honesty, we define a *type space* according to Bergemann and Morris (2005, 2012) and Matsushima and Noda (2020):

---

[7] We denote by $\Delta(Z)$ the space of probability measures on the Borel field of a measurable space Z. If Z is finite and $\rho \in \Delta(Z)$ satisfies $\rho(z) = 1$ for some $z \in Z$, I will simply write $\rho = z$.

[8] This study considers both deterministic SCFs and stochastic SCFs.

$$\Gamma \equiv (T_i, \pi_i, \theta_i)_{i \in N},$$

where $t_i \in T_i$ is agent $i's$ type, $\pi_i : \Omega \times T_i \to \Delta(T_{-i})$, and $\theta_i : \Omega \times T_i \to \{0,1\}$.[9] Agent i is selfish (honest) if $\theta_i(\omega, t_i) = 0$ ($\theta_i(\omega, t_i) = 1$, respectively). Agent i expects that the other agents' types are distributed according to the probability measure $\pi_i(\omega, t_i) \in \Delta(T_{-i})$.

3. Mechanism and Honesty

The central planner requires each agent to announce a message (or a bundle of multiple sub-messages) concerning the state simultaneously. To incentivize them to announce sincerely, the central planner designs a mechanism $G \equiv (M, g, x)$, where $M = \underset{i \in N}{\times} M_i$, M_i denotes agent $i's$ message space, $g : M \to \Delta(A)$ denotes an allocation rule, $x = (x_i)_{i \in N}$ denotes a payment rule, and $x_i : M \to [-\varepsilon, \varepsilon]$ denotes the payment rule for agent i. Here, $\varepsilon > 0$ is an arbitrary positive real number that implies limited solvency. Each agent i simultaneously announces a message $m_i \in M_i$, and the central planner determines the allocation according to $g(m) \in \Delta(A)$, paying the monetary transfer $x_i(m) \in R$ to each agent i. We assume budget balancing in that

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

We apply the method of bounded mechanism design[10] explored by Abreu and Matsushima (1992), which has been considered to play a decisive role in solving the unique implementation of SCFs. From a semantical point of view, we focus on the following class of mechanisms. We fix an arbitrary positive integer $K > 0$. (The specification of K will be subsequently explained). Let $M_i = \underset{k=0}{\overset{K}{\times}} M_i^k$ and

---

[9] We denote $Z \equiv \underset{i \in N}{\times} Z_i$, $Z_{-i} \equiv \underset{j \neq i}{\times} Z_j$, $z = (z_i)_{i \in N} \in Z$, and $z_{-i} = (z_j)_{j \neq i} \in Z_{-i}$.

[10] See Jackson (1992) for the definition of boundedness, which is a requirement on the plausibility of mechanisms. we assume the compactness of message spaces and continuity, which guarantees this boundedness.

$$M_i^k = \Delta(\Omega) \quad \text{for all} \quad k \in \{0,1,...,K\},$$

where we denote $m_i = (m_i^k)_{k=0}^K$ and $m_i^k \in M_i^k$ for each $k \in \{0,1,...,K\}$. With this specification, each agent i reports $K+1$ sub-messages at once, which typically concern which state actually occurs. At each k-th sub-message, agent i announces a distribution over states $m_i^k \in \Delta(\Omega)$. (An agent can announce different distributions across sub-messages). This specification serves to evoke an ethical motive from an agent in a manner such that the agent feels guilty about telling a lie that generates more material benefit.

A strategy for agent i is defined as

$$s_i : \Omega \times T_i \to M_i,$$

according to which agent i with type t_i announces $m_i = s_i(\omega, t_i) \in M_i$ in state ω. Let $s_i = (s_i^k)_{k=0}^K$, $s_i^k : \Omega \times T_i \to M_i^k$, and $s_i(\omega, t_i) = (s_i^k(\omega, t_i))_{k=0}^K$, where $s_i^k(\omega, t_i) \in M_i^k = \Delta(\Omega)$ denotes agent $i's$ k-th sub-message. We define the *sincere strategy* for agent i, denoted by $s_i^* = (s_i^{*k})_{k=0}^K$, as

$$s_i^{*k}(\omega, t_i) = \omega \quad \text{for all} \quad k \in \{0,...,K\},$$

according to which agent i announces about the state truthfully for any sub-message.

Each agent $i's$ material benefit is given by a quasi-linear utility $u_i(a, \omega) + r_i$, provided the central planner determines the allocation $a \in A$ and gives the monetary transfer $r_i \in R$ to agent i at the state $\omega \in \Omega$. If agent i is selfish and only concerned with the material benefit, the agent maximizes the expected value of material benefit as follows:

$$[\theta_i(\omega, t_i) = 0]$$
$$\Rightarrow [s_i(\omega, t_i) \in \arg\max_{m_i \in M_i} E[u_i(g(m), \omega) + x_i(m) \mid \omega, t_i, s_{-i}]],$$

where we assumed that the other agents announce according to $s_{-i} = (s_j)_{j \neq i}$.[11]

On the other hand, if agent i is honest, the agent is motivated not only by material benefit but also by an *intrinsic preference for honesty*. That is, an honest agent

[11] $E[\cdot \mid \xi]$ denotes the expectation operator conditional on ξ.

has a psychological cost $c_i(m,\omega,t_i,G)\in R$ such that for every $\omega\in\Omega$, $m\in M$, and $\tilde{m}_i\in M_i$,

(1) $[\theta_i(\omega,t_i)=1,\ m_i^{-k}=\tilde{m}_i^{-k}$, and $m_i^k(\omega)>\tilde{m}_i^k(\omega)]$

$\Rightarrow [c_i(\tilde{m}_i,m_{-i},\omega,t_i,G)\geq c_i(m,\omega,t_i,G)]$ for all $k\in\{0,...,K\}$,

and

(2) $[\theta_i(\omega,t_i)=1,\ m_i^{-0}=\tilde{m}_i^{-0},\ m_i^0(\omega)>\tilde{m}_i^0(\omega)$, and

$u_i(g(\tilde{m}_i,m_{-i}),x_i(\tilde{m}_i,m_{-i}),\omega)>u_i(g(m),x_i(m),\omega)]$

$\Rightarrow [c_i(\tilde{m}_i,m_{-i},\omega,t_i,G)>c_i(m,\omega,t_i,G)]$,

where we denoted $m_i^{-k}=(m_i^{k'})_{k'\neq k}$. An honest agent i maximizes the expected value of the material benefit minus the psychological cost as follows:

$$[\theta_i(\omega,t_i)=1]\Rightarrow [s_i(\omega,t_i)\in\arg\max_{m_i\in M_i} E[u_i(g(m),\omega)+x_i(m)$$

$$-c_i(m,\omega,t_i,G)\,|\,\omega,t_i,s_{-i}]].$$

From (1), any honest type weakly prefers announcing more honestly than the selfish types for any sub-message. From (2), any honest type strictly prefers announcing more honestly than the selfish types for the 0-th sub-message, provided the lie generates more material benefit. we allow the psychological cost to be arbitrarily small. This definition of psychological cost is an extension of that in Matsushima and Noda (2020) where each agent announces just a single message. We permit an honest agent to tell a white lie when it does not influence the agent's material benefit. This permission makes the results of this study robust against the possibility of other behavioral motives, provided that these motives are not as important as honesty.[12]

This study uses Bayes Nash equilibrium (BNE) as the solution concept in the game associated with a mechanism G. Each agent $i's$ state-contingent payoff function with type t_i in the game associated with a mechanism G, $U_i(\cdot;\omega,t_i):M\to R$, is defined as

$$U_i(m;\omega,t_i)=u_i(g(m),\omega)+x_i(m) \qquad \text{if } \theta_i(\omega,t_i)=0,$$

[12] Matsushima and Noda (2020) provide a more detailed discussion about other behavioral motives.

and

$$U_i(m;\omega,t_i) = u_i(g(m),\omega) + x_i(m) - c_i(m,\omega,t_i,G)$$
$$\text{if } \theta_i(\omega,t_i) = 1.$$

A strategy profile s is said to be a *Bayes Nash equilibrium* (BNE) in the game associated with the mechanism G if, for every $\omega \in \Omega$, $i \in N$, $t_i \in T_i$, and $m_i \in M_i$,

$$E[U_i(s_i(\omega,t_i),m_{-i};\omega,t_i,G) \mid \omega,t_i,s_{-i}]$$
$$\geq E[U_i(m_i,m_{-i};\omega,t_i,G) \mid \omega,t_i,s_{-i}].$$

4. Unique Implementation

A mechanism G is said to *uniquely implement* an SCF f if the sincere strategy profile s^* is the unique BNE, and it induces the value of f without monetary transfers, that is, for every $\omega \in \Omega$ and $t \in T$,

$$g(s^*(\omega,t)) = f(\omega) \text{ and } x(s^*(\omega,t)) = 0,$$

where we denote $s(\omega,t) = (s_i(\omega,t_i))_{i\in N}$. An SCF is said to be *uniquely implementable* if there exists a mechanism that uniquely implements it.

We demonstrate an epistemological condition, implying that "all agents are selfish" never happens to be common knowledge, as follows.[13] We call a subset of type profiles $E \subset T$ an event. Consider an arbitrary state $\omega \in \Omega$ and an arbitrary event $E \subset T$. Let

$$V_i^1(E,\omega) \equiv \{t_i \in T_i \mid \pi_i(E \mid \omega,t_i) = 1\},$$

and

$$V_i^k(E,\omega) \equiv \{t_i \in T_i \mid \pi_i(\underset{j\in N}{\times} V_j^{k-1}(E,\omega) \mid \omega,t_i) = 1\} \text{ for each } k \geq 2,$$

where we denote

$$E_{-i}(t_i) \equiv \{t_{-i} \in T_{-i} \mid (t_i,t_{-i}) \in E\} \text{ and } \pi_i(E \mid \omega,t_i) \equiv \pi_i(E_{-i}(t_i) \mid \omega,t_i).$$

[13] The following definitions are based on Matsushima and Noda (2020).

Here, $V_i^1(E,\omega)$ implies the set of agent $i's$ types with which agent i knows that the event E and the state ω occur, and $V_i^k(E,\omega)$ implies the set of agent $i's$ types with which agent i knows that the event $\underset{j\in N}{\times} V_j^{k-1}(E,\omega)$ and the state ω occur. We define

$$V_i^\infty(E,\omega) \equiv \bigcap_{k=1}^{\infty} V_i^k(E,\omega).$$

The event $E \subset T$ is said to be *common knowledge* at $(\omega,t) \in \Omega \times T$ if

$$t \in \underset{i\in N}{\times} V_i^\infty(E,\omega).$$

We denote by $E^*(\omega) \subset T$ the event that the state ω occurs and all agents are selfish, that is,

$$E^*(\omega) \equiv \{t \in T \mid \forall i \in N : \theta_i(\omega,t_i) = 0\}.$$

Theorem 1: *An SCF* f *is uniquely implementable if*

(3) $$\underset{i\in N}{\times} V_i^\infty(E^*(\omega),\omega) = \phi \quad \textit{for all} \quad \omega \in \Omega.$$

Condition (3) implies that "all agents are selfish" never happens to be common knowledge. Hence, Theorem 1 states that *any SCF is uniquely implementable if "all agents are selfish" never happens to be common knowledge.*

5. Mechanism Design

To prove Theorem 1, we design a mechanism G as follows. For each $k \in \{1,...,K\}$, we define $g^k : M^k \to \Delta(A)$ as a *majority rule* in the manner that for every $\omega \in \Omega$,

$g^k(m^k) = f(\omega)$ if $\frac{1}{n}\sum_{i\in N} m_i^k(\omega) > \frac{1}{n}\sum_{i\in N} m_i^k(\omega')$ for all $\omega' \neq \omega$, that is, the average of all agents' k-th sub-messages gives ω the highest probability,

and

$$g^k(m^k) = a^* \qquad \text{if there exists no such } \omega,$$

where $a^* \in A$ was selected arbitrarily. The central planner randomly selects $k \in \{1,...,K\}$ and determines the allocation according to $g^k(m^k) \in \Delta(A)$; hence, we specify the allocation rule g as follows:

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

Note that $g(m)$ is independent of the zero-th sub-message profile m^0.

For each $i \in N$ and $j \neq i$, we define $y_{i,j} : M_i^0 \times M_j^0 \to [-1,0]$ as a *quadratic scoring rule* (Brier, 1950):

$$y_{i,j}(m_i^0, m_j^0) = -\sum_{\omega \in \Omega} \{m_i^0(\omega) - m_j^0(\omega)\}^2,$$

which implies the distance between agent $i's$ 0-th sub-message and agent $j's$ 0-th sub-message. We denote by $r_i(m_i^{-0}, m_{i+1}^0) \in \{0,...,K\}$ the number of integers $k \in \{1,...,K\}$ such that $m_i^k \neq m_{i+1}^0$.[14] We define $w_i : M \to [-2,0]$ as follows:

$$w_i(m) = -\frac{r_i(m_i^{-0}, m_{i+1}^0)}{K} - 1$$

if there exists $k \in \{1,...,K\}$ such that $m_i^k \neq m_{i+1}^0$,

and $m_j^{k'} = m_{j+1}^0$ for all $k' < k$ and $j \in N$,

and

$$w_i(m) = -\frac{r_i(m_i^{-0}, m_{i+1}^0)}{K}$$

if there exists no such $k \in \{1,...,K\}$.

We then specify the payment rule x_i for agent i as follows:

$$x_i(m) = \frac{\varepsilon}{3}\left[\frac{1}{n-1}\left\{\sum_{j \neq i} y_{i,j}(m_i^0, m_j^0) - \frac{1}{n-2}\sum_{i' \neq i, j \neq i, i' \neq j} y_{i',j}(m_{i'}^0, m_j^0)\right\}\right.$$

$$\left. + w_i(m) - w_{i+1}(m)\right].$$

---

[14] I denote $i+1=1$ if $i=n$. I denote $m_i^{-0} = (m_i^1,...,m_i^K)$ and $m^{-0} = (m^k)_{k=1}^K$.

Note that the specified payment rule x satisfies budget balancing and limited solvency.

Let us select $K > 0$ sufficiently large to satisfy

$$K > \frac{3}{2\varepsilon} \max_{(a,a') \in A^2} \{v_i(a,\omega) - v_i(a',\omega)\}. \tag{4}$$

Note that the sincere strategy profile s^* satisfies that for every $(\omega,t) \in \Omega \times T$, $m \in M$, and $i \in N$,

$$g^k(m^k) = f(\omega) \qquad \text{if } m_{-i} = s^*_{-i}(\omega, t_{-i}),^{15}$$

$$x_i(m) = 0 \qquad \text{if } m = s^*(\omega,t),$$

and

$$x_i(m) < 0 \qquad \text{if } m_{-i} = s^*_{-i}(\omega, t_{-i}) \text{ and } m_i \neq s^*_i(\omega, t_i).$$

This implies that s^* is a BNE, and it achieves the value of the SCF f without monetary transfers.

The Appendix shows the proof that if a strategy profile s is a BNE, then $s = s^*$ must hold. The next section briefly explains the underlying logic.

6. Outline of the Proof

This section shows the outline of the proof that if s is a BNE, then $s = s^*$ must hold. The proof is divided into two parts: "*unique information elicitation*" and "*unique implementation with provability*," in the following manner.

Part 1 (Unique Information Elicitation): Part 1 corresponds to the finding in Matsushima and Noda (2020). Part 1 shows that $s^0 = s^{*0}$, that is, every agent whether selfish or honest announces the state truthfully for the 0-th sub-message. Note that each agent $i's$ 0-th sub-message influences her welfare only through the sum of the values of the quadratic scoring rules $\sum_{j \neq i} y_{i,j}(m_i^0, m_j^0)$ as well as the agent's psychological cost. Hence, from the nature of the quadratic scoring rule, any selfish type prefers mimicking

[15] I denote $s_{-i}(\omega, t_{-i}) = (s_j(\omega, t_j))_{j \neq i}$.

the average of the other agents' 0-th sub-messages. However, any honest type prefers announcing (slightly) more honestly than selfish types. These are the driving forces that tempt even selfish types to announce about the state truthfully for the 0-th sub-messages. According to Matsushima and Noda (2020), all agents, whether selfish or honest, announce about the state truthfully for their 0-th sub-messages as unique equilibrium behavior if "all agents are selfish" never happens to be common knowledge. Hence, $s^0 = s^{*0}$ must hold.

Part 2 (Unique Implementation with Provability): Assume $s^0 = s^{*0}$. Part 2 shows that $s^k = s^{*k}$, that is, all agents announce about the state truthfully for their k-th sub-messages, for all $k \in \{1, ..., K\}$. The designed mechanism implies that each agent i regards the neighbor's (i.e., agent $i+1$'s) 0-th sub-message as *reference*, and is tempted to announce this reference for any sub-message $k \in \{1, ..., K\}$. Part 1 shows that this reference equals the true state in equilibrium, that is, which state actually occurs was provable. This will tempt all agents to announce about the state truthfully for every sub-message.

To understand the logic behind Part 2, consider a case in which limited solvency ε is sufficiently large to satisfy

$$\frac{2\varepsilon}{3} > \max_{(a,a') \in A^2} \{v_i(a, \omega) - v_i(a', \omega)\}. \tag{5}$$

From (4) and (5), we can select $K = 1$, and therefore, simply write the designed mechanism as follows: for every $\omega \in \Omega$ and $m \in M$ such that $m_i^0 = \omega$ for all $i \in N$,

$g(m) = f(\tilde{\omega})$ if there exists $\tilde{\omega} \in \Omega$ such that

$$\frac{1}{n} \sum_{i \in N} m_i^1(\tilde{\omega}) > \frac{1}{n} \sum_{i \in N} m_i^1(\omega') \text{ for all } \omega' \neq \tilde{\omega},$$

$g(m) = a^*$ if there exists no such $\tilde{\omega}$,

and

$$x_i(s_i^*(\omega, t_i), m_{-i}) - x_i(m) = \frac{2\varepsilon}{3}$$

if $m_i^1 \neq s_i^{*1}(\omega, t_i) = \omega$.

From (5), we have

$$x_i(s_i^*(\omega,t_i),m_{-i})-x_i(m)>\max_{(a,a')\in A^2}\{v_i(a,\omega)-v_i(a',\omega)\}$$

$$\geq v_i(g(s_i^*(\omega,t),m_{-i}),\omega)-v_i(g(m),\omega).$$

Hence, the penalty on lying for the 1-st sub-message is greater than the impact of this lying on the determination of allocation. Hence, $s^1=s^{*1}$ must hold.

According to Abreu and Matsushima (1992), we can extend this observation to the general case where ε is an arbitrary positive real number and K is selected sufficiently large to satisfy (4). The designed mechanism incentivizes each agent to avoid being the first liar starting from the 1-st sub-message and also provides each agent i with a slight incentive to reduce the number $r_i(m_i^{-0},m_{i+1}^0)$. This method will drive all agents into a tail-chasing competition toward honest reporting across all sub-messages. Hence, $s^{-0}=s^{*-0}$ must hold.

Remark: This study proved the uniqueness of the pure strategy BNE. However, in the same manner as in Theorem 1, we can also prove the uniqueness of the *mixed strategy* BNE. In fact, in Part 1, any agent has the unique best response of the 0-th sub-message to any mixture of the other agents' 0-th sub-messages. In Part 2, we eliminated all unwanted strategies through the iterative dominance process. This guarantees the uniqueness of not only pure but also mixed strategy BNE.

7. Ethical SCF

This section assumes that $\theta_i(\omega,t_i)$ and $\pi_i(\omega,t_i)$ are independent of ω, and there exists a common prior over type profiles. Hence, $E^*(\omega)$ and $V_i^\infty(E,\omega)$ are independent of ω. We write E^* and $V_i^\infty(E)$ instead of $E^*(\omega)$ and $V_i^\infty(E,\omega)$, respectively.

An SCF f is said to be *ethical* if there exist two states that imply the same material utilities for all agents, but the SCF f assigns different allocations, that is, there exist $\omega\in\Omega$, $\omega'\in\Omega$, and $(q_i)_{i\in N}\in R^n$ such that

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

and

$$v_i(a,\omega) = v_i(a,\omega') + q_i \quad \text{for all} \quad i \in N \quad \text{and} \quad a \in A.$$

An SCF f is said to be *material* if it is not ethical, that is, for every $\omega \in \Omega$, $\omega' \in \Omega$, and $(q_i)_{i \in N} \in R^n$,

$$[\, v_i(a,\omega) = v_i(a,\omega') + q_i \quad \text{for all} \quad i \in N \quad \text{and} \quad a \in A \,]$$
$$\Rightarrow [\, f(\omega) = f(\omega') \,].$$

A material SCF depends only on the information about agents' material interests.

We show that any ethical SCF fails to be uniquely implementable if "all agents are selfish" happens to be common knowledge. That is, an ethical SCF is uniquely implementable if and only if "all agents are selfish" is never common knowledge, i.e., condition (3) holds.

Theorem 2: *If an SCF f is uniquely implementable and*

$$\underset{i \in N}{\times} V_i^{\infty}(E^*) \neq \phi,$$

then it is material.

Proof: Suppose that f is ethical and uniquely implemented by a mechanism G, where s denotes the unique BNE. Consider $(\omega, \omega') \in \Omega^2$ where $f(\omega) \neq f(\omega')$ and there exists $(q_i)_{i \in N} \in R^n$ such that $v_i(a,\omega) = v_i(a,\omega') + q_i$ for all $i \in N$ and $a \in A$. From the common prior assumption, we can specify a strategy profile $\tilde{s}$ as follows: for every $(\tilde{\omega}, t) \in \Omega \times T$,

$$\tilde{s}(\tilde{\omega}, t) = s(\tilde{\omega}, t) \qquad \text{if either} \quad t \notin \underset{i \in N}{\times} V_i^{\infty}(E^*) \quad \text{or} \quad \tilde{\omega} \neq \omega',$$

and

$$\tilde{s}(\tilde{\omega}, t) = s(\omega, t) \qquad \text{if} \quad t \in \underset{i \in N}{\times} V_i^{\infty}(E^*) \quad \text{and} \quad \tilde{\omega} = \omega'.$$

Clearly, $\tilde{s}$ is a BNE. However, this, along with the non-emptiness of $\underset{i \in N}{\times} V_i^{\infty}(E^*)$, contradicts the unique implementation because

$$g(\tilde{s}(\omega', t)) = f(\omega) \neq f(\omega') \quad \text{for all} \quad t \in \underset{i \in N}{\times} V_i^{\infty}(E^*).$$

Q.E.D.

A SCF f is said to be *uniquely and virtually implementable* if for every $\delta > 0$, there exists an SCF $\tilde{f}$ that is uniquely implementable and satisfies

$$\sum_{\omega \in \Omega} \left| f(\omega) - \tilde{f}(\omega) \right| \leq \delta .$$

We show that any material SCF is uniquely and virtually implementable irrespective of whether condition (3) holds, while any ethical SCF is uniquely and virtually implementable if and only if condition (3) holds.

Theorem 3: *Any material SCF is uniquely and virtually implementable. An ethical SCF is uniquely and virtually implementable if and only if*

$$\underset{i \in N}{\times} V_i^{\infty}(E^*) = \phi .$$

Proof: According to Abreu and Matsushima (1992), any material SCF is uniquely and virtually implementable whenever "all agents are selfish" is common knowledge, that is, $\underset{i \in N}{\times} V_i^{\infty}(E^*) = T$. Even without this assumption, the method of proof in Abreu and Matsushima is effective, because honest agents are more likely to be honest than selfish agents and rather work positively for iterative elimination of unwanted (dishonest) strategies. Hence, any material SCF is uniquely and virtually implementable in our epistemological framework.

Suppose that condition (3) does not hold, i.e., $\underset{i \in N}{\times} V_i^{\infty}(E^*) \neq \phi$. Consider an arbitrary ethical SCF f. Then, there exists $\delta > 0$ such that for any SCF $\tilde{f}$, whenever $\sum_{\omega \in \Omega} \left| f(\omega) - \tilde{f}(\omega) \right| \leq \delta$, then $\tilde{f}$ is ethical. Since Theorem 1 implies that any ethical SCF is not uniquely implementable, we have proved that f is not uniquely and virtually implementable.

Q.E.D.

8. Conclusion

This study investigated a society where people are either selfish or weakly honest, and showed that every SCF, whether material or ethical, is uniquely implementable in BNE if "all agents are selfish" never happens to be common knowledge. A material SCF is always uniquely and virtually implementable, while an ethical SCF is uniquely and virtually implementable if and only if "all agents are selfish" never happens to be common knowledge.

We assumed symmetric information concerning the state across agents. Although not specifically shown in this study, according to the same manner as Matsushima and Noda (2020), we can weaken this assumption and then extend the analysis to a case of asymmetric information where each agent does not necessarily access the (full) information channel and does not even know who are actually informed.

It is an important future research to investigate the unique implementation of ethical SCFs in a more general asymmetric information environment, where each agent can observe only partial information about the state as private information. Matsushima (2008a) showed a permissive result in this environment by assuming that all agents never tell white lies: Matsushima (2008a) modified the construction for Part 2 of this study by changing each agent's reference from her neighbor's 0-th sub-message to her own early-stage sub-message. By using the same modification as Matsushima (2008a), we can show that even with weak honesty, any SCF is uniquely implementable in the general environment if (and only if) the central planner can solve the corresponding information elicitation problem. However, it is an open question to clarify under what condition the central planner can solve this problem.[16] Answering this question is by no means easy, and goes beyond the scope of this study.

[16] Matsushima and Noda (2020) provided an example in which the central planner can solve the information elicitation problem even in the asymmetric information environment.

References

Abeler, J., D. Nosenzo, and C. Raymond (2019): Preference for Truth-Telling, Econometrica 87 (4), 1115-1153.

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

Arrow, K. (1951): *Social Choice and Individual Values*, Yale University Press.

Bergemann, D., and S. Morris (2005): Robust mechanism design, Econometrica 73, 1771-1813.

Bergemann, D. and S. Morris (2012): An Introduction to Robust Mechanism Design, Foundations and Trends in Microeconomics 8 (3), 169-230.

Brier, G. (1950): Verification of Forecasts Expressed in Terms of Probability, Monthly Weather Review, 78, 1-3.

Dogan, B. (2017): Eliciting the Socially Optimal Allocation from Responsible Agents, Journal of Mathematical Economics 73, 103-110.

Dutta, B. and A. Sen (2012): Nash Implementation with Partially Honest Individuals, Games and Economic Behavior 74 (1), 154-169.

Gibbard, A. (1973): Manipulation of Voting Schemes: A General Result, Econometrica 41 (4), 587-601.

Hurwicz, L. (1972): On Informationally Decentralized Systems, in Decision and Organization, ed. by C.B. McGuire and R. Radner. North Holland, Amsterdam.

Jackson, M. (2001): A Crash Course in Implementation Theory, Social Choice and Welfare 18, 655-708.

Kartik, N., O. Tercieux, and R. Holden (2014): Simple Mechanisms and Preferences for Honesty, Games and Economic Behavior 83, 284-290.

Lombardi, M. and N. Yoshihara (2017): Natural Implementation with Semi-Responsible Agents in Pure Exchange Economies, International Journal of Game Theory 46 (4), 1015-1036.

Lombardi, M. and N. Yoshihara (2018): Treading a Fine Line: (Im) possibilities for Nash Implementation with Partially-Honest Individuals, Games and Economic Behavior 111, 203-216.

Lombardi, M. and N. Yoshihara (2019): Partially-Honest Nash Implementation: A Full

Characterization, Economic Theory 54 (1), 1-34.

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

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. (2008a): Role of Honesty in Full Implementation, Journal of Economic Theory 139 (1), 353-359.

Matsushima, H. (2008b): Behavioral Aspects of Implementation Theory, Economics Letters 100 (1), 161-164.

Matsushima, H. (2013): Process Manipulation in Unique Implementation, Social Choice and Welfare 41 (4), 883-893.

Matsushima, H. (2019): Implementation without Expected Utility: Ex-Post Verifiability, Social Choice and Welfare 53 (4), 575-585.

Matsushima, H. and S. Noda (2020): "Epistemological Mechanism Design," CARF-F-498, UTMD-001, University of Tokyo.

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.

Mukherjee, S., N. Muto, and E. Ramaekers (2017): Implementation in Undominated Strategies with Partially Honest Agents, Games and Economic Behavior 104, 613-631.

Ortner, J. (2015): Direct Implementation with Minimally Honest Individuals, Games and Economic Behavior 90, 1-16.

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.

Saporiti, A. (2014): Securely Implementable Social Choice Rules with Partially Honest Agents, Journal of Economic Theory 154, 216-228.

Satterthwaite, M. (1975): Strategy-Proofness and Arrow's Conditions: Existence and Correspondence Theorems for Voting Procedures and Social Welfare Functions, Journal of Economic Theory 10 (2), 187-217.

Savva, F. (2018): Strong Implementation with Partially Honest Individuals, Journal of

Mathematical Economics 78, 27-34.

Yadav, S. (2016): Selecting Winners with Partially Honest Jurors, Mathematical Social Sciences 83, 35-43.

Appendix: Proof of Theorem 1

Suppose that a strategy profile s is a BNE. I fix an arbitrary state $\omega \in \Omega$. First, we show that

$$s_i^0(\omega, t_i) = \omega \quad \text{for all} \quad i \in N \quad \text{and} \quad t_i \in T_i .$$

Since the selection of m_i^0 influences agent $i's$ welfare only through the sum of the values of the quadratic scoring rules $\sum_{j \neq i} y_{i,j}(m_i^0, m_j^0)$ and the psychological cost, the following properties are obtained:

$$[\theta_i(\omega, t_i) = 0]$$
$$\Rightarrow [s_i^0(\omega, t_i) \in \arg\max_{m_i \in M_i} E[\sum_{j \neq i} y_{i,j}(m_i^0, m_j^0) \mid \omega, t_i, s_{-i}, G]],$$

and

$$[\theta_i(\omega, t_i) = 1]$$
$$\Rightarrow [s_i(\omega, t_i) \in \arg\max_{m_i \in M_i} E[\sum_{j \neq i} y_{i,j}(m_i^0, m_j^0) - c_i(m_i, \omega, t_i) \mid \omega, t_i, s_{-i}, G] .$$

From the nature of the quadratic scoring rule, we can calculate the best response as follows:

$$[\theta_i(\omega, t_i) = 0] \Rightarrow [s_i^0(\omega, t_i) = E[\frac{\sum_{j \neq i} s_i^0(\omega, t_j)}{n-1} \mid \omega, t_i]],$$

and

$$[\theta_i(\omega, t_i) = 1] \Rightarrow [\text{either} \quad s_i^0(\omega, t_i)(\omega) = 1 \quad \text{or}$$
$$s_i^0(\omega, t_i)(\omega) > E[\frac{\sum_{j \neq i} s_i^0(\omega, t_j)(\omega)}{n-1} \mid \omega, t_i]].$$

That is, any selfish agent mimics the average of the other agents' zero-th sub-messages in expectation, while any honest agent announces more honestly than the selfish types. This will drive agents into a tail-chasing competition, reaching the point at which all agents report honestly for their zero-th sub-messages. We can directly apply the theorem in Matsushima and Noda (2020) to this situation, and we therefore can prove

that any BNE satisfies $s_i^0(\omega, t_i) = \omega$ as long as the equalities (3) hold: $s_i^0 = s_i^{*0}$ must hold for all $i \in N$.

Second, I prove that

$$s_i^k(\omega, t_i) = \omega \quad \text{for all } k \in \{1, ..., K\},\ i \in N, \text{ and } t_i \in T_i.$$

The specification of x_i implies that if an agent i announces a sub-message different from the neighbor's (agent $(i+1)'s$) zero-th sub-message as the first deviation starting from the 1-st sub-message, she is fined the monetary amount $\varepsilon/3$. Since I have selected K sufficiently large, that is, the inequality (4) holds, the impact of the selection of each sub-message on the determination of the allocation is sufficiently small compared with the monetary amount $\varepsilon/3$. Hence, the mechanism design for this theorem is based on the method explored by Abreu and Matsushima (1992), the so-called A-M mechanisms. This will drive agents into another tail-chasing competition through which each agent avoids becoming the first deviant. Since we have already proved that all agents announce truthfully for their zero-th sub-messages ($m_i^0 = \omega$ for all $i \in N$), this competition drives them to announce the state truthfully for all sub-messages.

To be precise, let us consider an arbitrary $k \in \{1, ..., K\}$, and suppose that $s^{k'} = s^{*k'}$ for all $k' < k$. If $m_j^k \neq \omega$ for some $j \neq i$, agent i strictly prefers announcing truthfully for the k-th sub-message because she can avoid being the first deviant. Even if $m_j^k = \omega$ for all $j \neq i$, agent i still strictly prefers announcing truthfully for the k-th sub-message because she does not want to increase $r_i(m_i^{-0}, m_{i+1}^0)$. Hence, through the iterative elimination of dominated strategies, we can inductively prove that

$$s_i^k = s_i^{*k} \quad \text{for all } i \in N \text{ and } k \in \{1, ..., K\}.$$

That is, there exists no BNE other than the sincere strategy profile s^*.

Since s^* is a BNE and achieves the value of f without monetary transfers we have completed the proof of Theorem 1.