UTMD Working Paper
The University of Tokyo
Market Design Center

UTMD-013

Epistemological Implementation of Social Choice Functions

Hitoshi Matsushima
University of Tokyo

July 20, 2021

Epistemological Implementation of Social Choice Functions[1]

Hitoshi Matsushima[2]

University of Tokyo

July 20, 2021

Abstract

We investigate the implementation of social choice functions (SCFs) from an epistemological perspective. We consider the possibility that in higher-order beliefs there exists an honest agent who is motivated by intrinsic preference for honesty as well as material interest. We assume weak honesty, in that an honest agent is mostly motivated by material interests and even tells white lies. Importantly, this study assumes that "all agents are selfish" never happens to be common knowledge. We then show the following positive results for the implementability: In complete information environments, with three or more agents, any SCF is uniquely implementable in the Bayesian Nash equilibrium (BNE). In asymmetric information environments, with a minor restriction named information diversity, any incentive-compatible SCF is fully implementable in BNE. An SCF, whether material or nonmaterial (ethical), can be implemented even if all agents are selfish and "all agents are selfish" is mutual knowledge.

Keywords: implementation, weak honesty, common knowledge on selfishness, ethical social choice function, information diversity
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 Japanese Ministry of Education, Culture, Sports, Science, and Technology (MEXT). An earlier version of this study is Matsushima (2020). In the present version of this study, part of the joint work by Hitoshi Matsushima and Shunya Noda (Epistemological Mechanism Design, Discussion Paper UTMD-001, University of Tokyo, 2020, Theorem 1) has been integrated into Section 4. Moreover, the analysis of asymmetric information environments (Section 5) is a theoretical contribution that has been newly added to this version. I am grateful to Professor Shunya Noda for his 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 implementation problem of social choice functions (SCFs) from an epistemological perspective. A central planner attempts to implement the desirable allocation implied by an SCF contingent on the state. She (or he)[3] does not know the state, while there exist multiple agents (participants) who are informed of it, fully (complete information) or partly (asymmetric information). The central planner attempts to incentivize these agents to announce 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 as a unique equilibrium outcome?

We assume that each agent is either selfish or honest. A selfish agent is only concerned about her 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. Instead, we consider the epistemological possibility that an honest agent exists, not in the mechanism, but in the participants' higher-order beliefs. By considering the epistemological type space, we show that a slight possibility of an honest agent in higher-order beliefs incentivizes all agents, whether selfish or honest, to behave sincerely.

This study assumes that "all agents are selfish (i.e., all agents are motivated only by their monetary interests)" never happens to be common knowledge. With this, we demonstrate the following positive results for the implementability: In complete information environments, with three or more agents, any SCF is uniquely implementable in the Bayesian Nash equilibrium (BNE). In asymmetric information environments, any incentive-compatible SCF is fully implementable in BNE whenever the private signal structure satisfies information diversity; that is, any observation of a private signal is informative in that the resultant posterior is different from the prior, and the informativeness of a private signal is diversified in that no private signal

[3] For convenience, this study uses the gendered pronoun "she," instead of "she or he" or "they."

changes the prior in the same direction as any other private signal does. Information diversity is a very weak restriction, because the posterior distribution does not necessarily reveal an agent's private signal.

With complete information, we utilize a simple form of the quadratic scoring rule (Brier, 1950), which aligns agents' payoffs with the distance between their messages, as a part of the payment rule design. The quadratic scoring rule plays a significant role in incentivizing all agents, whether selfish or honest, to announce sincerely as unique BNE behavior, whenever the common knowledge of selfishness is eliminated.

The usefulness of the simple form of the quadratic scoring rule, however, crucially depends on the assumption of complete information. Hence, for asymmetric information environments, we create a new, more elaborate, method of the quadratic scoring rule design, where we do not require the sincere strategy profile to be the unique BNE. Instead, we require each agent to gradually reveal her private signal through multiple announcements.

This study requires only a weak honesty. Even honest agents are motivated mostly by their monetary interests; that is, the influences of the intrinsic preferences for honesty on decision making are arbitrarily small. Agents do not expect the possibility that there exists an honest participant; that is, they have mutual knowledge that all agents are selfish (i.e., all agents know that all agents are selfish). Despite these weaknesses in honesty, the central planner can elicit correct information from all agents if "all agents are selfish" never happens to be common knowledge.

2. Literature Review

The early literature on social choice and implementation theory has assumed that "all agents are selfish" is common knowledge, and focused on those considerations of SCFs that are material, that is, those that depend only on agents' material utilities (Arrow, 1951; Hurwicz, 1972; Gibbard, 1973; Satterthwaite, 1975; Maskin, 1977/1999; Abreu and Matsushima, 1992a; 1992b).[4] With this common knowledge, it is

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

impossible in principle to implement any SCF that is nonmaterial; that is, it depends not only on an agent's material utilities but also on nonmaterial factors such as ethics, equity, fairness, future generation, and environmental concerns. This study suggests a highly positive potential for implementing such nonmaterial SCFs.[5]

Matsushima (2008a) and Matsushima (2008b) are the pioneering works incorporating an intrinsic preference for honesty into implementation theory and demonstrating a new research trend to overcome the above-mentioned difficulty. Matsushima (2008a) showed that in complete information environments, with three (or more) agents, any SCF, whether material or nonmaterial, is uniquely implementable if there exists a weakly honest agent who dislikes telling a white lie. Matsushima (2008b) showed that in asymmetric information environments, any incentive-compatible SCF, whether material or nonmaterial, is uniquely implementable if all agents are honest. Many subsequent studies such as Dutta and Sen (2012), Kartik, Tercieux, and Holden (2014), Saporiti (2014), Ortner (2015), and Mukherjee, Muto, and Ramaekers (2017) have studied the complete information environments and showed their respective positive results.[6]

This study makes two significant advances in this line of research under complete information. 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 only rule out the case in which "all agents are selfish" is common knowledge: we permit the case where all agents are selfish and "all agents are selfish" is mutual knowledge.

Second, previous works assumed that an honest agent never tells a white lie, that is, a lie that does not influence material utilities. However, given that real people may be more or less influenced by various irrational motives, we contend that this assumption is restrictive. In contrast to this assumption, this study permits even honest

[5] An exception is Matsushima (2019; 2021a), which assumed that the state is ex-post verifiable and proved that any SCF, whether material or nonmaterial, is uniquely implementable even if "all agents are selfish" is common knowledge.

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

agents to tell white lies: an honest agent feels guilty about lying only when this lying increases the agent's material benefit.

Many empirical and experimental studies indicate that human beings are not purely motivated by monetary payoffs but have an intrinsic preferences for honesty. Abeler, Nosenzo, and Raymond (2019) provided a detailed meta-analysis using data from 90 studies involving more than 44,000 subjects across 47 countries, showing that subjects who were in trade-offs between material interest and honesty gave up a large fraction of potential benefits from lying.[7]

We permit that an honest agent can tell white lies. We do not assume that there exists an honest agent in the mechanism. To make full use of the slight possibility of weak honesty in epistemology, we utilize the quadratic scoring rules, which can set aside various non-selfish motives and prioritize the agent's monetary interest. As Abeler, Nosenzo, and Raymond point out, the intrinsic preference for honesty remains included, and an honest agent prefers announcing 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 agents, reaching a point at which all agents report honestly in the complete information environment.

The quadratic scoring rule is one of the standard mechanism design methods for partial implementation.[8] However, if "all agents are selfish" is common knowledge, there exists a serious multiplicity of unwanted BNEs: "all agents tell the same lie" is a BNE, regardless of what this lie is.

[7] Various works in behavioral economics and decision theory have modeled preferences for honesty, such as a cost of lying (e.g., Ellingsen and Johannesson, 2004; Kartik, 2009; Kartik et al, 2007; Kartik and Tercieux, 2012), a reputational cost (e.g., Mazar, Amir, and Ariely, 2008), and guilt aversion (e.g., Charness and Dufwenberg, 2006).

[8] See Cooke (1991) for a survey of scoring rules. For the applications to mechanism design, see for example Johnson et al. (1990), Matsushima (1990; 1991; 1993; 2007), Aoyagi (1998), and Miller et al. (2007). A number of studies extended the scoring rule of Brier (1950) to a setting in which a central planner collects information from a group of agents (e.g., Dasgupta and Ghosh, 2013; Prelec, 2004; Miller et al., 2005; Kong and Schoenebeck, 2019). Previous studies commonly assumed that all agents are selfish and, thus, suffered from the multiplicity of equilibria in a "single-question" setting in which the state is realized only once (as in our model).

Matsushima and Noda (2020) first found in the context of information elicitations that truth-telling is a unique BNE if there exists an honest agent, thereby suggesting that the quadratic scoring rule design is a potentially powerful solution not only for partial implementation but also for unique implementation. This study generalizes the usefulness of the quadratic scoring rule design by showing positive results for the general implementability of SCFs.

Research on asymmetric information environments has made little progress since Matsushima (2008b), although that study's sufficient condition was quite restrictive and more careful analyses were eagerly awaited. The second half of this study is devoted to this consideration, and demonstrates a sufficient condition for the full implementation named "information diversity," which is a very weak restriction on the asymmetric information structure. Information diversity implies that any observation of a private signal is informative in the sense that the resultant posterior is different from the prior, and that the informativeness of a private signal is diversified in the sense that no private signal changes the prior in the same direction as any other private signal does. Information diversity does not require the distribution to reveal an agent's type. Hence, information diversity is much weaker than any informational restriction discussed in the implementation literature such as Bayesian monotonicity (Jackson, 1991), measurability (Abreu and Matsushima, 1992b), or no consistent deception (Matsushima, 1993). With information diversity and without common knowledge on selfishness, this study demonstrates a new design of the quadratic scoring rules and proves that the full implementation of all incentive-compatible SCFs is possible.

The equilibrium analysis of games with behavioral agents and asymmetric information has a long history. Kreps et al. (1982) studied how the existence of behavioral agents changes the equilibria of finitely repeated games. Postlewaite and Vives (1987), Carlsson and van Damme (1993), and Morris and Shin (1998) studied how incomplete information shrinks the set of equilibria. These studies focused on the analysis of given games, while our focus is on the design of mechanisms that can take full advantage of the potential existence of behavioral agents in higher-order beliefs.

Similar to Rubinstein's (1989) email game, this study contrasts the outcome under common knowledge and "almost common knowledge." Rubinstein's (1989) email game demonstrates that "almost common knowledge" could lead to an unintuitive outcome; while our study demonstrates the vulnerability of the common knowledge assumption, its implication contrasts Rubinstein's. The intuitive outcome is truth-telling, and people can naturally expect that a truthful strategy profile is a focal point, while there are many equilibria under common knowledge of selfishness. By carefully designing a "game" (mechanism), we can eliminate all the unwanted and unintuitive equilibria where "all agents are selfish" is not common knowledge (while it could be "almost common knowledge").

The remainder of this paper is organized as follows. The basic model is presented in Section 3. Section 4 considers the complete information environments in which all agents are fully informed of the state. We semantically define a class of indirect mechanisms and then define the intrinsic preference for honesty. We define the unique implementation in BNE and state that any SCF is uniquely implementable if "all agents are selfish" never happens to be common knowledge (Theorem 1). Section 5 considers the asymmetric information environments in which each agent is informed of the state only partly. We demonstrate information diversity, define the unique implementation in BNE, and then show that with information diversity, any incentive-compatible SCF is fully implementable in BNE if "all agents are selfish" never happens to be common knowledge (Theorem 2). Section 6 concludes.

3. The Model

This study investigates a situation in which a central planner attempts to achieve a desirable allocation contingent on the state. Let $N \equiv \{1,...,n\}$ denote the finite set of all agents, where $n \geq 2$. Let A denote the non-empty and finite set of all the allocations. Let Ω denote the non-empty and finite set of states. The *social choice function* (SCF) is defined as $f:\Omega \rightarrow \Delta(A)$.[9] For every $\omega \in \Omega$, $f(\omega) \in \Delta(A)$ implies a desirable distribution of allocation at state ω .[10] We assume that the central

[9] 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 $\rho = z$.

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

planner does not know the state, while all agents are informed of the state fully (Section 4) or partly (Section 5).

Each agent is either *selfish* or *honest*. Details on the meaning of selfishness and honesty will be explained in Sections 4 and 5. No agent knows if the other agents are selfish or honest. To describe agents' higher-order beliefs concerning their selfishness and honesty, we define an *epistemological type space* as follows:

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

where $t_i \in T_i$ is agent $i's$ epistemological type, $\pi_i : T_i \to \Delta(T_{-i})$, and $\theta_i : T_i \to \{0,1\}$.[11] Agent i is selfish (honest) if $\theta_i(t_i) = 0$ ($\theta_i(t_i) = 1$, respectively). Agent i expects that the epistemological types of other agents are distributed according to the probability measure $\pi_i(t_i) \in \Delta(T_{-i})$. We assume that there exists a common prior $\pi \in \Delta(\Omega)$ from which $(\pi_i)_{i\in N}$ is derived.

We call a subset of epistemological type profiles $E \subset T \equiv \times_{i\in N} T_i$ an event. For convenience, for each event $E \subset T$, we write $\pi_i(E \,|\, t_i) = \pi_i(E_{-i}(t_i) \,|\, t_i)$, where we denote $E_{-i}(t_i) \equiv \{t_{-i} \in T_{-i} \,|\, (t_i, t_{-i}) \in E\}$. We denote by $E^* \subset T$ the event that all agents are selfish, that is,

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

Consider an arbitrary event $E \subset T$. Let

$$V_i^1(E) = \{t_i \in T_i \,|\, \pi_i(E \,|\, t_i) = 1\}.$$

Recursively, for each positive integer $h \geq 2$, let

$$V_i^h(E) = \{t_i \in T_i \,|\, \pi_i(\underset{j\in N}{\times} V_j^{h-1}(E) \,|\, t_i) = 1\}.$$

We then define

$$V_i^\infty(E) \equiv \bigcap_{h=1}^{\infty} V_i^h(E).$$

An event $E \subset T$ is said to be *common knowledge* at an epistemological type profile $t \in T$ if

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

[11] 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 $z_{-i} = (z_j)_{j\neq i} \in Z_{-i}$.

Note that if E is common knowledge at $t \in T$, then

$$\pi_i\left(\underset{j \in N}{\times} V_j^\infty(E) \middle| t_i\right) = 1 \text{ for all } i \in N.$$

Fix an arbitrary positive real number $\varepsilon > 0$. We define a mechanism as $G \equiv (M, g, x)$, 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, $\varepsilon > 0$ implies the level of limited liability. Each agent i simultaneously announces a message $m_i \in M_i$, and the central planner determines the allocation according to $g(m) \in \Delta(A)$ and pays the monetary transfer $x_i(m) \in R$ to each agent i.

Each agent $i's$ material benefit is given by a quasi-linear utility $v_i(a, \omega) + r_i$, provided that the central planner determines the allocation $a \in A$ and gives the monetary transfer $r_i \in R$ to agent i at state $\omega \in \Omega$. We assume expected utility for convenience, and denote $v_i(\alpha, \omega)$ the expected payoff derived from stochastic allocation $\alpha \in \Delta(A)$. When all agents announce $m \in M$ in the mechanism G, the resultant expected material payoff is given by $v_i(f(m), \omega) + x_i(m)$.

This study considers a small liability ε, which is positive but close to zero. Quasi-linearity is a convenient, but rather redundant, assumption: all we need for this study is that an agent's material benefit increases as the monetary transfer to her increases. The expected utility assumption is also redundant: see Matsushima (2019, 2021a) for this detail.

4. Complete Information

This section considers the complete information environments concerning the state, where all agents are fully informed of the state. We assume $n \geq 3$. From a semantic point of view, we focus on the following class of mechanisms. We fix an arbitrary positive integer $K \geq 1$, the specification of which is explained in Subsection 4.2. Let

$$M_i = \mathop{\times}_{k=1}^{K} M_i^k \text{ , and}$$

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

where we denote $m_i = (m_i^k)_{k=1}^K$, and $m_i^k \in M_i^k$ for each $k \in \{1,...,K\}$. Each agent i reports K sub-messages at once, which typically concern which state actually occurs. At each k-th sub-message, agent i announces a distribution over states $m_i^k \in \Delta(\Omega)$.

Under complete information, we define a strategy for agent i as

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

according to which, agent i with epistemological type t_i announces $m_i = s_i(\omega, t_i) \in M_i$ at the state ω. Denote $s_i = (s_i^k)_{k=1}^K$, $s_i^k : \Omega \times T_i \to M_i^k$, and $s_i(\omega, t_i) = (s_i^k(\omega, t_i))_{k=1}^K$, 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 $s_i^* = (s_i^{*k})_{k=1}^K$, as

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

according to which, agent i announces the state truthfully at any sub-message.

Each agent $i \in N$ is either selfish ($\theta_i(t_i) = 0$) or honest ($\theta_i(t_i) = 1$). If agent i is selfish, she is only concerned with the material benefit; that is, she maximizes the expected value of material benefit:

$$[\theta_i(t_i) = 0]$$

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

where we assumed that the other agents announce according to $s_{-i} = (s_j)_{j \neq i}$.[12]

If agent i is honest, she is motivated not only by material benefit but also by an *intrinsic preference for honesty*: she 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) $$[\theta_i(t_i) = 1,\ m_i(\omega) \neq \tilde{m}_i(\omega),\ m_i(\omega) \geq \tilde{m}_i(\omega), \text{ and}$$
$$v_i(g(\tilde{m}_i, m_{-i}), \omega) + x_i(\tilde{m}_i, m_{-i}) > v_i(g(m), \omega) + x_i(m)]$$

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

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

and

(2) $$[\theta_i(t_i) = 1 \text{ and } m_i(\omega) = \tilde{m}_i(\omega)]$$
$$\Rightarrow [c_i(\tilde{m}_i, m_{-i}, \omega, t_i, G) = c_i(m, \omega, t_i, G)],$$

where we denote $m_i(\omega) = (m_i^k(\omega))_{k=1}^K$. From (1), an honest agent feels guilty if she gains material payoffs from telling a lie. We do not impose any restriction on the strength of the psychological cost. She maximizes the expected value of the material benefit minus the psychological cost:

$$[\theta_i(t_i) = 1] \Rightarrow [s_i(\omega, t_i) \in \arg\max_{m_i \in M_i} E[v_i(g(m), \omega) + x_i(m)$$
$$-c_i(m, \omega, t_i, G) \mid \omega, t_i, s_{-i}]].$$

Formally, each agent $i's$ payoff function, $U_i(\cdot; \omega, t_i): M \to R$, is defined as

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

and

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

A strategy profile s is said to be a *Bayesian 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 $m_i \in M_i$,

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

A mechanism G is said to *uniquely implement* an SCF f if there exists the unique BNE s, and it induces the value of f; that is,

$$g(s(\omega, t)) = f(\omega) \text{ for all } \omega \in \Omega \text{ and } t \in T,$$

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.

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

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

Equality (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. The proof of Theorem 1 is presented in the subsequent subsections.

4.1. Special Case: Information Elicitation

To understand the proof of Theorem 1, it is helpful to investigate a special case called information elicitation, where each agent's material payoff is irrelevant to the allocation; that is,

$$v_i(a,\omega)=0 \text{ for all } i\in N,\ a\in A, \text{ and } \omega\in\Omega .$$

Proposition 1: *In the information elicitation problem, any SCF* f *is uniquely implementable if equality (3) holds.*

4.1.1. Mechanism Design

To prove Proposition 1, we design the following mechanism: $G=(M,g,x)$. Let $K=1$. Let

$$M_i=M_i^1=\Delta(\Omega) \text{ for all } i\in N,$$

and

$$g(m)=a^* \text{ for all } m\in M,$$

where a^* is selected arbitrarily. For each $i\in N$ and $j\neq i$, we specify $y_{i,j}:M_i^1\times M_j^1\to[-1,0]$ as a quadratic scoring rule:

$$y_{i,j}(m_i^1,m_j^1)=-\sum_{\omega\in\Omega}\{m_i^1(\omega)-m_j^1(\omega)\}^2 ,$$

which implies the distance between agent $i's$ 1-st sub-message and agent $j's$ 1-st sub-message. For each $i\in N$, we specify the payment rule: for every $m\in M$,

$$\hat{x}_i(m) = \frac{\varepsilon}{n-1}\sum_{j\neq i} y_{i,j}(m_i^1, m_j^1)$$

$$= -\frac{\varepsilon}{n-1}\sum_{j\neq i}[\sum_{\omega\in\Omega}\{m_i^1(\omega) - m_j^1(\omega)\}^2],$$

where we denote $m_i = m_i^1$ because of $K = 1$.

From the nature of the quadratic scoring rule, any selfish type prefers to mimic the average of the other agents' messages. Importantly, any honest type prefers announcing slightly more honestly than selfish types. These are the driving forces that tempt even selfish types to announce truthfully.

4.1.2. Proof of Proposition 1

From the nature of the quadratic scoring rules, it is clear that the sincere strategy profile s^* is a BNE in the game associated with the specified mechanism G; thus, it suffices to show uniqueness.

Suppose that s is a BNE. Fix $\omega\in\Omega$ arbitrarily. Let

$$\alpha = \min_{(i,t_i)} s_i(\omega, t_i)(\omega),$$

and

$$\tilde{T}_i \equiv \{t_i \in T_i \mid s_i(\omega, t_i)(\omega) = \alpha\} \text{ for each } i \in N.$$

Suppose that Eq. (3) holds. Note that Eq. (3) is equivalent to

$$V_i^\infty(E^*) = \phi \text{ for all } i\in N.$$

Suppose that $\alpha < 1$, that is, there exists an agent $i\in N$ with type $t_i \in T_i$ that does not adopt the sincere strategy. Note from the definition of psychological cost that any honest agent prefers making announcements more honestly than selfish agents. Hence, no honest type belongs to $\tilde{T}_i$; that is, $\tilde{T}_i \subset E_i^*$.

Let us consider an arbitrary $i\in N$ and $t_i \in \tilde{T}_i$. Note that α equals the average of the other agents' announcements on ω in expectation but not greater than any announcement. Hence, an epistemological type t_i assumes that any other agent $j\neq i$ announces $m_i(\omega) = \alpha$, that is,

$$\pi_i\left(\underset{j\in N}{\times}\tilde{T}_j\middle|t_i\right)=1.$$

This, along with the definition of the psychological cost, implies that agents i with epistemological type t_i expect that the other agents are surely selfish, that is,

$$\pi_i(E^*\mid t_i)=1.$$

Hence, we have

$$\tilde{T}_i\subset V_i^1(E^*).$$

Moreover, since

$$\pi_i\left(\underset{j\in N}{\times}V_j^1(E^*)\middle|t_i\right)\geq\pi_i\left(\underset{j\in N}{\times}\tilde{T}_j\middle|t_i\right)=1,$$

we have $\pi_i(\times_{j\in N}V_j^1(E^*)\mid t_i)=1$, that is,

$$\tilde{T}_i\subset V_i^2(E^*).$$

Similarly, we have

$$\tilde{T}_i\subset V_i^k(E^*)\quad\text{for all}\quad k\geq 2.$$

Hence, we have

$$\tilde{T}_i\subset V_i^\infty(E^*),$$

which contradicts the assumption that $V_i^\infty(E^*)=\phi$. Hence, we conclude that $\alpha=1$, or, equivalently, $s_i(\omega,t_i)=\omega$ for all $\omega\in\Omega$. Accordingly, $s=s^*$ must be the case for any BNE. From these observations, we prove Proposition 1.

Remark 1: It is well known that quadratic scoring rules incentivize agents to be honest as a BNE in the information elicitation problem. The seminal work of Matsushima and Noda (2020) first pointed out that truth-telling is not only a BNE but also a unique NBE, provided that "all agents are selfish" is not mutual knowledge. Proposition 1 applies the logic of higher-order beliefs such as email games (Rubinstein,1989) and global games (Carlsson and van Damme, 1993; Morris and Shin, 1998) to the information elicitation problem, and succeeded in extending the findings of Matsushima and Noda (2020) to the situation in which "all agents are selfish" may be mutual knowledge but is not common knowledge.

4.2. General Case

4.2.1. Mechanism Design

To prove Theorem 1 generally, we design another mechanism G as follows. Let $K \geq 3$. Let

$$M_i^1 = \Delta(\Omega), \text{ and}$$

$$M_i^k = \Omega \text{ for all } k \in \{2,...,K\}.$$

For each $k \in \{3,...,K\}$, we specify $g^k : M^k \to \Delta(A)$ as a majority rule: for every $\omega \in \Omega$,

$$g^k(m^k) = f(\omega) \quad \text{if } m_i^k = \omega \text{ for more than } {}^{n}\!/_{2} \text{ agents,}$$

and

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

The central planner randomly selects $k \in \{3,...,K\}$ and determines the allocation according to $g^k(m^k) \in \Delta(A)$; that is, we specify the allocation rule g:

$$g(m) = \frac{\sum_{k=3}^{K} g^k(m^k)}{K-2} \quad \text{for all } m \in M.$$

It is important to note that $g(m)$ is independent of the 1-st and 2-nd sub-messages, m^1 and m^2.

To specify the payment rule, we use the quadratic scoring rules $y_{i,j}$ as well as the following functions, z_i, w_i, and r_i; that is, we define $z_i : M_i^2 \times M_{i+1}^1 \to [-1,0]$:

$$z_i(m_i^2, m_{i+1}^1) = -1 \quad \text{if } m_i^2 \neq m_{i+1}^1,$$

and

$$z_i(m_i^2, m_{i+1}^1) = 0 \quad \text{if } m_i^2 = m_{i+1}^1,\text{[13]}$$

which implies that the agent i is fined if her 2-nd sub-message is different from her neighbor's 1-st sub-message. We define $w_i : M \to [-1,0]$:

[13] We denote $i+1=1$ if $i=n$.

$$w_i(m) = -1 \qquad \text{if there exists } k \in \{3,...,K\} \text{ such that } m_i^k \neq m_i^2 \text{, and } m_j^{k'} = m_j^2 \text{ for all } k' < k \text{ and } j \in N,$$

and

$$w_i(m) = 0 \qquad \text{if there exists no such } k \in \{3,...,K\},$$

which implies that agent i is fined if she is the first deviant from the own 2-nd sub-message. We further define $r_i(m_i) \in \{0,...,K-2\}$ as the number of integers $k \in \{3,...,K\}$ such that $m_i^k \neq m_i^2$, that is, the number of agent $i's$ sub-messages after her 3-rd sub-message that are different from her 2-nd sub-message.

Fix an arbitrary positive real number $\xi > 0$, which is set sufficiently large. We specify the payment rule x_i for agent i:

$$x_i(m) = \frac{\varepsilon}{3+\xi}\{\frac{1}{n-1}\sum_{j\neq i} y_{i,j}(m_i^1, m_j^1) + \xi z_i(m_i^2, m_{i+1}^1)$$
$$+ w_i(m) - \frac{r_i(m_i)}{K-2}\}.$$

Note that the specified payment rule x satisfies limited solvency; that is,

$$x_i(m) \in [-\varepsilon, \varepsilon] \text{ for all } i \in N \text{ and } m \in M.$$

Let us select $K \geq 3$ sufficiently large to satisfy

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

With $n \geq 3$, the sincere strategy profile s^* satisfies that for every $(\omega, t) \in \Omega \times T$, $m \in M$, and $i \in N$,

$$g^k(m^k) = f(\omega) \qquad \text{if } m_{-i} = s_{-i}^*(\omega, t_{-i}),$$ [14]

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

and

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

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

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

The next subsection will show that if a strategy profile s is a BNE in the game associated with the specified mechanism, then $s = s^*$ must hold, which completes the proof of Theorem 1.

4.2.2. Proof of Theorem 1

The proof of Theorem 1 is divided into two parts: "*information elicitation*" and "*implementation with provability*" in the following manner.

Part 1 (Information Elicitation): Part 1 shows that $s^1 = s^{*1}$, that is, every agent, whether selfish or honest, announces the state truthfully for the 1-st sub-message. Note that each agent $i's$ 1-st sub-message influences her welfare only through $\sum_{j \neq i} y_{i,j}(m_i^1, m_j^1)$. Hence, we can directly apply Proposition 1 and show that $s^1 = s^{*1}$.

Part 2 (Implementation with Provability): Assume that a BNE strategy profile s satisfies $s^1 = s^{*1}$. Part 2 shows that $s^k = s^{*k}$ for all $k \in \{2, ..., K\}$, that is, all agents announce the state truthfully at their remaining sub-messages.

First, because $\xi z_i(m_i^2, m_{i+1}^1)$ imposes a relatively large fine, each agent i is willing to announce truthfully for the 2-nd sub-message, irrespective of the other sub-messages of this agent.

Each agent i regards her 2-nd sub-message as reference, and she is tempted to announce this reference at any sub-message $k \in \{3, ..., K\}$ of this agent. Given that this reference is equivalent to the true state in equilibrium, that is, the state that actually occurs is substantially provable, it follows that all agents are tempted to announce the state truthfully at every sub-message.

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

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

From (4) and (5), we can select $K=3$ and simply write the designed mechanism as follows: for every $\omega\in\Omega$ and $m\in M$ such that $m_i^1 = m_i^2 = \omega$ for all $i\in N$,

$$g(m) = f(\tilde{\omega}) \quad \text{if } m_i^3 = \omega \text{ for more than } n/2 \text{ agents,}$$

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

and

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

$$\text{if } m_i^3 \neq s_i^{*3}(\omega,t_i) = \omega.$$

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 3-rd sub-message is greater than the impact of this on the determination of allocation, and $s^3 = s^{*3}$ must hold.

Following Abreu and Matsushima (1992a), we can extend this observation to the case where ε is small, if we select a sufficiently large K to satisfy (4). The designed mechanism incentivizes each agent to avoid being the first deviant starting from the 3-rd sub-message and also provides each agent i with an incentive to reduce the number $r_i(m_i)$. This method drives all agents into a tail-chasing competition toward honest reporting from the 3-rd to the K-th sub-messages. Hence, $s = s^*$ must hold.

The formal proof is as follows. Suppose that s is a BNE. Fix an arbitrary state $\omega\in\Omega$. First, we show that

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

Because the selection of m_i^1 influences agent $i's$ welfare only through $\sum_{j\neq i} y_{i,j}(m_i^1,m_j^1)$ and the psychological cost, the following properties are obtained:

$$[\theta_i(t_i) = 0]$$

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

and

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

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

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

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

and

$$[\,\theta_i(t_i)=1\,] \Rightarrow [\text{either } \; s_i^1(\omega,t_i)(\omega)=1 \; \text{ or}$$

$$s_i^1(\omega,t_i)(\omega) > E[\frac{\sum_{j\neq i} s_i^1(\omega,t_j)(\omega)}{n-1} \,|\, \omega,t_i\,]\,].$$

In other words, any selfish agent mimics the average of the other agents' 1-st 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 at their 1-st sub-messages. Hence, from Proposition 1, we can prove that any BNE s satisfies $s_i^1 = s_i^{*1}$ for all $i \in N$.

Because $\xi z_i(m_i^2, m_{i+1}^1)$ imposes a relatively large fine, it follows from $s_{i+1}^1 = s_{i+1}^{*1}$ that each agent i is willing to select $m_i^2 = \omega$. Hence, $s_i^2 = s_i^{*2}$ must hold for all $i \in N$.

We further prove that

$$s_i^k(\omega,t_i) = \omega \; \text{ for all } \; k \in \{3,...,K\}, \; i \in N \text{ , and } \; t_i \in T_i\,.$$

The specifications of w_i and x_i imply that if an agent i announces a sub-message different from her 2-nd sub-message as the first deviation starting from the 3-rd sub-message, she is fined the monetary amount $\frac{\varepsilon}{3+\xi}$. Because we have selected a sufficiently large K, that is, 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 $\frac{\varepsilon}{3+\xi}$. This will drive agents into another tail-chasing competition, through which each agent avoids becoming the first deviant. Because we have already proved that all agents announce truthfully at their 2-nd sub-messages, this competition drives them to announce the state truthfully at all sub-messages.

To be precise, consider an arbitrary $k \in \{3,...,K\}$ 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 at 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 at the k-th sub-message because she does not want to increase $r_i(m_i)$. Hence, through the iterative elimination of dominated strategies, we can inductively prove that $s_i^k = s_i^{*k}$ for all $i \in N$ and $k \in \{3,...,K\}$. In other words, there exists no BNE other than the sincere strategy profile s^*.

Because s^* is a BNE and achieves the value of f, we have completed the proof of Theorem 1.

Remark 2: To better understand mechanism G, let us consider the following decision procedure: The central planner asks each agent to input a distribution on Ω. The central planner also asks each agent to input an element of Ω, and gives all agents a one-hour grace period. During this continuous period, each agent can change her second input at any time and number of times. After this grace period, the first and second inputs become public. The central planner selects one point from the grace period and then determines the allocation according to the majority rule and their inputs at this point. The central planner imposes just a small monetary fine to any agent who is the last person to change the second input. The central planner imposes another small, but slightly larger, monetary fine to any agent whose initial second input is different from her neighbor’s first input. The central planner also makes monetary transfers according to the quadratic scoring rule and the first inputs.

Because of the nature of the quadratic scoring rule, all agents are willing to make their first inputs equivalent to the true state. Given that the possibility of selecting a point in the continuous period is negligible, any agent prefers to make her initial input equal to the true state and avoid becoming the last person to change the second input: she is willing to keep her correct input unchanged during the grace period.

Remark 3: In the proof of Theorem 1, we have proved the uniqueness of not only the pure but also the mixed-strategy BNE. In fact, in Part 1, any agent has the unique best response of the 1-st sub-message to any mixture of the other agents' 1-st sub-messages. In Part 2, we eliminated all unwanted strategies through an iterative dominance process. This guarantees the uniqueness of the mixed-strategy BNE.

Remark 4: Without substantial difficulty, we can replace the payment rule with a budget-balancing payment rule. We redefine $\boldsymbol{w}_i$ by replacing the first deviant among all agents with the first deviant among all agents other than some agent and set it as the transfer from this agent. We also set other parts of the payment rule as the transfers from agents whose announcements are irrelevant. This modification can be applied to the remaining arguments in this study (Section 5).

5. Asymmetric Information

This section considers asymmetric information environments, where each agent is informed of the state only partly. Let

$$\boldsymbol{\Omega} = \underset{i\in N}{\times} \boldsymbol{\Omega}_i,\ \ \boldsymbol{\omega}_i \in \boldsymbol{\Omega}_i, \text{ and } \boldsymbol{\omega} = (\boldsymbol{\omega}_i)_{i\in N} \in \underset{i\in N}{\times} \boldsymbol{\Omega}_i .$$

Each agent $\boldsymbol{i} \in \boldsymbol{N}$ is informed of the $\boldsymbol{i-th}$ component $\boldsymbol{\omega}_i$ as a private signal (material type). We denote by $\boldsymbol{p}_{i,j}(\cdot \,|\, \boldsymbol{\omega}_i) : \boldsymbol{\Omega}_j \to [\mathbf{0},\mathbf{1}]$ the probability distribution on $\boldsymbol{\Omega}_j$ which is conditional on $\boldsymbol{\omega}_i \in \boldsymbol{\Omega}_i$. For each $\boldsymbol{\alpha}_i \in \Delta(\boldsymbol{\Omega}_i)$, let $\boldsymbol{p}_{i,j}(\cdot \,|\, \boldsymbol{\alpha}_i) = \sum_{\omega_i \in \Omega_i} \boldsymbol{p}_{i,j}(\cdot \,|\, \boldsymbol{\omega}_i)\boldsymbol{\alpha}_i(\boldsymbol{\omega}_i)$. We assume that there exists a common prior $\boldsymbol{p} : \boldsymbol{\Omega} \to [\mathbf{0},\mathbf{1}]$ from which $\boldsymbol{p}_{i,j}(\cdot \,|\, \boldsymbol{\omega}_i)$ is derived. Let $\boldsymbol{p}_i : \boldsymbol{\Omega}_i \to [\mathbf{0},\mathbf{1}]$ denote the prior distribution over $\boldsymbol{\Omega}_i$, where

$$p_i(\omega_i) = \sum_{\omega_{-i} \in \Omega_{-i}} p(\omega) \text{ for all } \omega_i \in \Omega_i .$$

We assume that the epistemological type space T is finite, and that ω and t are independently drawn.

We assume the following condition on the prior p, which is a very weak restriction:

Information Diversity: For every $i \in N$, $j \in N \setminus \{i\}$, $\omega_i \in \Omega_i$, and $\omega_i' \neq \omega_i$, there exists no $\beta \geq 0$ such that

$$p_{i,j}(\cdot \mid \omega_i) - p_j(\cdot) = \beta \left\{ p_{i,j}(\cdot \mid \omega_i') - p_j(\cdot) \right\} .$$

Information diversity implies that any observation of a private signal is informative in the sense that the resultant posterior is different from the prior, and that the informativeness of a private signal is diversified in the sense that no private signal changes the prior in the same direction as any other private signal does. Information diversity is a very weak restriction because the distribution does not necessarily reveal an agent's type. In fact, for each $\omega_i \in \Omega_i$, information diversity permits the existence of a mixture $\alpha_i \in \Delta(\Omega_i) \setminus \{\omega_i\}$ such that $p_{i,j}(\cdot \mid \alpha_i) = p_{i,j}(\cdot \mid \omega_i)$.

From a semantic point of view, this section considers a class of mechanisms (M, g, x) in which there exists a positive integer L such that

$$M_i = \mathop{\times}_{l=1}^{L} M_i^l \text{, and}$$

$$M_i^l \subset \Delta(\Omega_i) \text{ for all } l \in \{1, ..., L\} .$$

Each agent i announces L sub-messages at once, concerning the distribution on Ω_i.

An agent is either selfish ($\theta_i(t_i) = 0$) or honest ($\theta_i(t_i) = 1$). An honest agent i has a psychological cost $c_i(m, \omega, t_i, G) \in R$ such that for every $\omega \in \Omega$, $m \in M$, and $\tilde{m}_i \in M_i$,

$$[\theta_i(t_i) = 1,\ m_i(\omega_i) \neq \tilde{m}_i(\omega_i),\ m_i(\omega_i) \geq \tilde{m}_i(\omega_i) \text{, and}$$

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

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

and

$$[\,\theta_i(t_i) = 1 \ \text{ and } \ m_i(\omega_i) = \tilde{m}_i(\omega_i)\,]$$

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

where we denote $m_i(\omega_i) = (m_i^l(\omega_i))_{l=1}^{L}$. In the same manner as in the complete information environment, an honest agent feels guilty if she gains material payoffs from telling a lie. We do not impose any restriction on the strength of the psychological cost.

Each agent $i's$ payoff function is defined as

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

and

$$U_i(m; \omega, t_i) = v_i(g(m), \omega) + x_i(m) - c_i(m, \omega, t_i, G)$$

$$\text{if } \theta_i(t_i) = 1.$$

The strategy for agent i is defined as $s_i : \Omega_i \times T_i \to M_i$. A strategy profile s is said to be a *Bayesian Nash equilibrium* (BNE) in the game associated with the mechanism G if for every $i \in N$, $\omega_i \in \Omega_i$, $t_i \in T_i$, and $m_i \in M_i$,

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

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

A mechanism G is said to *fully implement* an SCF f if there exists a BNE, and any BNE s satisfies

$$g(s(\omega, t)) = f(\omega) \ \text{ for all } \ \omega \in \Omega \ \text{ and } \ t \in T,$$

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

An SCF f is said to be *incentive compatible* if for every $i \in N$ and $\omega_i \in \Omega_i$,

$$E[v_i(f(\omega)) \mid \omega_i] \geq E[v_i(f(\omega_i', \omega_{-i})) \mid \omega_i] \ \text{ for all } \ \omega_i' \in \Omega_i.$$

Theorem 2: *Any incentive-compatible SCF is fully implementable if information diversity holds and "all agents are selfish" never happens to be common knowledge; that is, Eq. (3) holds.*

The proof of Theorem 2 will be shown in the subsequent subsections.

5.1. Special Case: Information Elicitation

To understand the proof of Theorem 2, it is helpful to investigate the special case called information elicitation, where we assume that

$$v_i(a,\omega)=0 \quad \text{for all } i\in N\,,\ a\in A\,, \text{and } \omega\in\Omega\,.$$

Proposition 2: *In the information elicitation problem, any SCF* f *is fully implementable if information diversity and Eq. (3) hold.*

5.1.1. Mechanism Design

In the complete information environment, we used the simplest form of the quadratic scoring rule for the unique implementation, where the sincere strategy profile was considered as the unique BNE. However, we cannot directly apply this method to the asymmetric information environment, because the method crucially depends on the common knowledge of the state among the agents. Because of this application failure, we will create a new, more elaborate, quadratic scoring rule design, in which each agent is required to make multiple announcements in the information elicitation problem. This design does not incentivize each agent to be honest. Instead, it incentivizes each agent to announce partial information about her private signal at a sub-message.

To prove Proposition 2, we design the following mechanism $G=(M,g,x)$: Fix an arbitrary positive integer H. Let

$$L=(n-1)H\,.$$

We denote (j,h) and $M_{i,j}^h$ instead of l and M_i^l. Let

$$M_i = \mathop{\times}_{h=1}^{H} M_i^h,$$

$$M_i^h = \mathop{\times}_{j\neq i} M_{i,j}^h,$$

and $M_{i,j}^h$ is specified as a subset of $\Delta(\Omega_i)$:

$$M_{i,j}^h = \{\alpha_i \in \Delta(\Omega_i) \,|\, \exists(\omega_i,\lambda)\in\Omega_i\times[0,1]: \alpha_i = \lambda\omega_i + (1-\lambda)p_i\}.$$

Each agent i simultaneously announces multiple $(n-1)H$ sub-messages. For convenience, we call $m_i^h = (m_{i,j}^h)_{j\neq i} \in M_i^h \equiv \mathop{\times}_{j\neq i} M_{i,j}^h$ the h-th sub-message, and $m_{i,j}^h \in M_{i,j}^h$ its sub-sub-message.

From information diversity, for each $m_{i,j}^h \in M_{i,j}^h \setminus \{\alpha_i\}$, there exists the unique (ω_i,λ) such that

$$m_{i,j}^h = \lambda\omega_i + (1-\lambda)p_i.$$

Hence, we can define $I_i : \Delta(\Omega_i) \to \Delta(\Omega_i)$ as follows:

$$I_i(\alpha_i) = \omega_i \qquad \text{if } \alpha_i = \lambda\omega_i + (1-\lambda)p_i \text{ for some } \lambda > 0,$$

and

$$I_{i,j}(\alpha_i) = p_i \qquad \text{if there exists no such } \omega_i.$$

Hence, $I_i(\alpha_i) = \omega_i$ implies that the announcement of α_i reveals private signal ω_i.

For each $h \in \{1,...,H\}$, we define $I_i^h : M_i^h \to \Delta(\Omega_i)$ as follows:

$$I_i^h(m_i^h) = \omega_i \qquad \text{if } I_i(m_{i,j}^h) = \omega_i \text{ for some } j \neq i \text{ and}$$
$$I_i(m_{i,j}^h) \in \{\omega_i, p_i\} \text{ for all } j \neq i,$$

and

$$I_i^h(m_i^h) = p_i \qquad \text{if there exists no such } \omega_i.$$

Hence, $I_i^h(m_i^h) = \omega_i$ implies that there exists an h-th sub-sub-message that reveals ω_i and there exists no h-th sub-sub-message of agent i that reveals a different private signal. In this case, agent i is considered to reveal ω_i in her h-th sub-message $m_i^h = (m_{i,j}^h)_{j\neq i}$.

We specify the allocation rule g: for every $m \in M$,

$$g(m) = f(\omega) \qquad \text{if } I_i^H(s_i^H(\omega_i, t_i)) = \omega_i \text{ for all } i \in N,$$

and

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

The central planner selects the allocation $f(\omega)$ if all agents' H-th announcements m^H reveal the state ω.

To specify the payment rule x, we define a quadratic scoring rule $\gamma_i : \Delta(\Omega_i)^2 \rightarrow R$:

$$\gamma_i(\alpha_i, \alpha_i') = -\sum_{\omega_i \in \Omega_i} \{\alpha_i(\omega_i) - \alpha_i'(\omega_i)\}^2.$$

Note that $\alpha_i = \alpha_i'$ uniquely maximizes $\gamma_i(\alpha_i, \alpha_i')$. Note that this equivalence holds even if α_i' is uncertain: the expected value of $\gamma_i(\alpha_i, \alpha_i')$ in terms of α_i' is uniquely maximized by selecting α_i equal to the expectation of α_i'. We define another quadratic scoring rule $\gamma_{i,j} : \Delta(\Omega_i) \times \Delta(\Omega_j) \rightarrow R$:

$$\gamma_{i,j}(\alpha_i, \alpha_j) = \gamma_j(p_{i,j}(\cdot \mid \alpha_i), \alpha_j).$$

If $p_{i,j}(\cdot \mid \alpha_i) = \alpha_j(\cdot)$, then α_i maximizes $\gamma_{i,j}(\alpha_i, \alpha_j)$. We then specify the payment rule x: for every $m \in M$ and $i \in N$,

$$x_i(m) = \frac{\varepsilon}{(n-1)H} \sum_{j \neq i} \left[\gamma_i(m_{i,j}^1, p_i) + \sum_{h=2}^{H} \gamma_{i,j}(m_{i,j}^h, I_j^{h-1}(m_j^{h-1})) \right].$$

Note that this specification satisfies the limited solvency.

5.1.2. Proof of Proposition 2

We show that the mechanism (M, g, x) designed in Subsection 5.1.1 fully implements the SCF f in the information elicitation problem. For each $h \in \{1, ..., H\}$, we define

$$T_i^h(\omega_i, s_i) \equiv \{t_i \in T_i \mid I_i(s_i^h(\omega_i, t_i)) = p_i\},$$

and

$$T_i^h(\omega_i, \omega_i', s_i) \equiv \{t_i \in T_i \mid I_i(s_i^h(\omega_i, t_i)) = \omega_i'\}.$$

According to the iterative eliminations of dominated strategies from the 1-st sub-messages to the H-th sub-messages, we prove that there exists a BNE, and any BNE s satisfies

$$T_i^H(\omega_i, \omega_i, s_i) = T_i \text{ for all } i \in N \text{ and } \omega_i \in \Omega_i .$$

Hence, the H-th sub-messages of all agents succeed in revealing the state fully through $(I_i)_{i\in N}$, regardless of the BNEs.

First, we consider the 1-st sub-messages. If agent i is selfish, she maximizes the quadratic scoring rule $\sum_{j\neq i}\gamma_i(m_{i,j}^1, p_i)$, which uniquely determines $s_{i,j}^1(\omega_i, t_i) = p_i$ for all $j \neq i$. Hence, we have

$$I_i^1(s_i^1(\omega_i, t_i)) = p_i .$$

If she is honest, she is willing to be more honest than the selfish types. In other words, she maximizes the quadratic scoring rule $\sum_{j\neq i}\gamma_i(m_{i,j}^1, p_i)$ minus her psychological cost, where because of the definition of the psychological cost, the selected $m_i^1 = s_i^1(\omega_i, t_i)$ satisfies that for every $j \neq i$, there exists $\lambda > 0$ such that $s_i^1(\omega_i, t_i) = \lambda\omega_i + (1-\lambda)p_i$. Hence, this honest agent truthfully reveals her private signal, that is,

$$I_i^1(s_i^1(\omega_i, t_i)) = \omega_i .$$

Note that an honest agent may have multiple best responses in this case. However, any best response m_i^1 satisfies $I_i^1(m_i^1) = \omega_i$. Accordingly, we have

$$T_i^1(\omega_i, s_i) \subset E^* ,$$

$$T_i^1(\omega_i, \omega_i, s_i) = T_i \setminus T_i^1(\omega_i, s_i) ,$$

$$\text{and } T_i^1(\omega_i, \omega_i', s_i) = \phi \text{ for all } \omega_i' \neq \omega_i .$$

Because $T_i^1(\omega_i, s_i)$ is independent of ω_i, we can write

$$T_i^1(s_i) = T_i^1(\omega_i, s_i) .$$

Next, consider the 2-nd sub-messages. If agent i is selfish and expects agent $j \neq i$ to belong to $T_j^1(s_j)$ with certainty, she maximizes the expected value of the second quadratic scoring rule $\gamma_{i,j}(m_{i,j}^2, I_j^1(m_j^1))$; that is, she maximizes $\gamma_{i,j}(m_{i,j}^2, p_j)$, and this maximization uniquely determines $s_{i,j}^2(\omega_i, t_i) = p_i$. Hence, we have

$$I_i^2(s_i^2(\omega_i, t_i)) = p_i .$$

If she is selfish and expects agent j to belong to $T_j \backslash T_j^1(s_j)$ with a positive probability, there exists $\lambda > 0$ such that she maximizes the expected value of $\gamma_{i,j}(m_{i,j}^2, \lambda\omega_j + (1-\lambda)p_j)$, and this maximization uniquely determines $s_i^2(\omega_i, t_i) = \lambda\omega_i + (1-\lambda)p_i$. Hence, this selfish agent reveals her private signal correctly; that is,we have

$$I_i^2(s_i^2(\omega_i, t_i)) = \omega_i .$$

If she is honest, then there exist $\lambda \geq 0$ and $\lambda' > 0$ such that she maximizes the expected value of $\gamma_{i,j}(m_{i,j}^2, \lambda\omega_j + (1-\lambda)p_j)$ minus her psychological cost, and this maximization selects $s_i^2(\omega_i, t_i) = \lambda'\omega_i + (1-\lambda')p_i$. Hence, this honest agent reveals her private signal correctly; that is, we have

$$I_i^2(s_i^2(\omega_i, t_i)) = \omega_i .$$

Accordingly, we have

$$T_i^2(\omega_i, s_i) \subset V_i^1(E^*) ,$$

$$T_i^2(\omega_i, \omega_i, s_i) = T_i \setminus T_i^2(\omega_i, s_i) ,$$

and

$$T_i^2(\omega_i, \omega_i', s_i) = \phi \text{ for all } \omega_i' \neq \omega_i .$$

Because $T_i^2(\omega_i, s_i)$ is independent of ω_i, we can write $T_i^2(s_i) = T_i^2(\omega_i, s_i)$.

Third, consider an arbitrary $h \in \{3, ..., H\}$ and the h-th sub-messages. Suppose that for every $i \in N$, $\omega_i \in \Omega_i$, and $h' \in \{1, ..., h-1\}$, $T_i^{h'}(\omega_i, s_i)$ is independent of ω_i, that is,

$$T_i^{h'}(s_i) \equiv T_i^{h'}(\omega_i, s_i) ,$$

$$T_i^{h'}(\omega_i, \omega_i, s_i) = T_i \setminus T_i^{h'}(s_i) ,$$

$$T_i^{h'}(\omega_i, \omega_i', s_i) = \phi \text{ for all } \omega_i' \neq \omega_i ,$$

and

$$T_i^{h'}(s_i) \subset V_i^{h'-1}(E^*).$$

In the same manner as the argument for the 2-nd sub-messages, if agent i is selfish and expects all agents $j \neq i$ to belong to $T_j^{h-1}(s_j)$ with certainty, then

$$I_i^h(s_i^h(\omega_i, t_i)) = p_i.$$

If she is selfish and expects some agent $j \neq i$ to belong to $T_j \backslash T_j^{h-1}(s_j)$ with a positive probability, then

$$I_i^h(s_i^h(\omega_i, t_i)) = \omega_i.$$

If she is honest, then

$$I_i^h(s_i^h(\omega_i, t_i)) = \omega_i.$$

Accordingly, we have

$$T_i^h(\omega_i, s_i) \subset V_i^{h-1}(E^*),$$

$$T_i^h(\omega_i, \omega_i, s_i) = T_i \setminus T_i^h(\omega_i, s_i),$$

$$T_i^h(\omega_i, \omega_i', s_i) = \phi \text{ for all } \omega_i' \neq \omega_i,$$

and $T_i^h(\omega_i, s_i)$ is independent of ω_i. We can write $T_i^h(s_i) = T_i^h(\omega_i, s_i)$.

From the above observations, we have

$$T_i^H(s_i) \subset V_i^{H-1}(E^*),$$

$$T_i^H(\omega_i, \omega_i, s_i) = T_i \setminus T_i^H(s_i),$$

and

$$T_i^H(\omega_i, \omega_i', s_i) = \phi \text{ for all } \omega_i' \neq \omega_i.$$

Because the epistemological type space is finite, we can derive the common knowledge event through finite steps of iterations; that is, for each event $E \subset T$, there exists a positive integer K such that

$$V_i^\infty(E) = V_i^k(E) \text{ for all } k \geq K \text{ and } i \in N.$$

From information diversity and Eq. (3) ($V_i^\infty(E^*) = \phi$), there exists K such that

$$V_i^k(E^*) = \phi \text{ for all } k \leq K.$$

By selecting $H > K$, we have

$$T_i^H(s_i) \subset V_i^{H-1}(E^*) = \phi,$$

which implies

$$T_i^H(\omega_i, \omega_i, s_i) = T_i,$$

that is,

$$I_i^H(s_i^H(\omega_i, t_i)) = \omega_i \text{ for all } i \in N, \ \omega_i \in \Omega_i, \text{ and } t_i \in T_i.$$

Hence, we have proved Proposition 2.

Unlike in complete information environments, here the central planner fails to incentivize each agent i to tell the truth literally. However, the central planner can obtain the correct information through indication I_i. In the 1-st sub-message, the central planner can obtain the correct information only from honest types through I_i, while she interprets selfish types as uninformative. In the 2-nd sub-message, the central planner can obtain the correct information from a wider range of types including selfish types, because, thanks to the second quadratic scoring rule, a selfish agent i is willing to make $m_{i,j}^2(\omega_i)$ greater than $p_i(\omega_i)$, that is, to reveal the true private signal through I_i, whenever she expects the other agents to reveal the true signals in the 1-st sub-message. By repeating the same reasoning, the central planer can get the correct information from more selfish types in later sub-messages. With the finiteness of the epistemological type space, with a sufficiently large H, and without common knowledge of all agents' selfishness, we can prove that the central planner can eventually obtain the correct information from all agents and all types at the H-th (final) sub-messages.

5.2. General Case

5.2.1. Mechanism Design

To prove Theorem 2, we design, as follows, a mechanism $G = (M, g, x)$, which is an extension of the mechanism designed in Subsection 5.1.1, where the manner of this extension is basically the same as in the complete information environments (Subsection 4.2.1). Fix arbitrary positive integers H and K, which are sufficiently large. Let

$$L = (n-1)H + K.$$

We specify

$$M_i = \overset{H+K}{\underset{k=1}{\times}} M_i^k .$$

For each $h \in \{1,...,H\}$, let

$$M_i^h = \underset{j \neq i}{\times} M_{i,j}^h ,$$

and $M_{i,j}^h$ is specified in the same manner as in Subsection 5.1.1:

$$M_{i,j}^h = \{\alpha_i \in \Delta(\Omega_i) \,|\, \exists(\omega_i, \lambda) \in \Omega_i \times [0,1] : \alpha_i = \lambda\omega_i + (1-\lambda)p_i\} .$$

For each $k \in \{H+1,...,H+K\}$, we specify

$$M_i^k = \Omega_i .$$

Each agent announces the h-th sub-message for each $h \in \{1,...,H\}$ in the same manner as the mechanism designed in the information elicitation problem (in Subsection 5.1.1). For each $k \in \{H+1,...,H+K\}$, she further announces an element of Ω_i as the k-th sub-message.

We specify the allocation rule g as follows:

$$g(m) = \frac{\sum_{k=H+2}^{H+K} f(m^k)}{K-1} \quad \text{for all } m \in M .$$

Note that $g(m)$ is independent of the first $H+1$ sub-messages $(m^h)_{h=1}^{H+1}$. The central planner randomly selects $k \in \{H+2,...,H+K\}$, and then determines the allocation according to $f(m^k) \in \Delta(A)$.

We specify the payment rule x as follows. We define $\hat{w}_i : M \to [-2,0]$ as follows:

$$\hat{w}_i(m) = -1 \qquad \text{if there exists } k \in \{H+2,...,H+K\} \text{ such that } m_i^k \neq m_i^{H+1}, \text{ and } m_j^{k'} = m_j^{H+1} \text{ for all } k' \in \{H+2,...,k-1\} \text{ and } j \in N ,$$

and

$$\hat{w}_i(m) = 0 \qquad \text{if there exists no such } k \in \{H+2,...,H+K\} .$$

Note that $\hat{w}_i(m)$ indicates whether agent i is the first deviant from the (H+1)-th sub-message. We denote by $\hat{r}_i(m_i) \in \{0,...,K-1\}$ the number of integers $k \in \{H+2,...,H+K\}$ such that $m_i^k \neq m_i^{H+1}$.

We specify the payment rule x_i for agent i as a combination of the payment rule specified in the information elicitation problem (in Subsection 4.1.1), the above-specified $\hat{w}_i$ and $\hat{r}_i$, and a quadratic score given by $\gamma_{i,j}(m_{i,j}^{H+1}, I_j^H(m_j^H))$; that is, for every $m \in M$ and $i \in N$,

$$x_i(m) = \frac{\varepsilon}{3+\xi}\left[\frac{1}{(n-1)H}\sum_{j\neq i}\left\{\gamma_i(m_{i,j}^1, p_i) + \sum_{h=2}^{H}\gamma_{i,j}(m_{i,j}^h, I_j^{h-1}(m_j^{h-1}))\right\}\right.$$
$$\left. +\xi\gamma_{i,j}(m_{i,j}^{H+1}, I_j^H(m_j^H)) + \hat{w}_i(m) - \frac{\hat{r}_i(m_i)}{K-1}\right],$$

where $\xi > 0$ is an arbitrarily positive real number that is set sufficiently large. Note that the specified x satisfies limited solvency. We select a sufficiently large H. We also select K sufficiently large to satisfy for every $i \in N$ and $\omega \in \Omega$,

(6) $$K > \frac{3+\xi}{\varepsilon}\max_{(a,a')\in A^2}\{v_i(a,\omega) - v_i(a',\omega)\} + 1.$$

5.2.2. Proof of Theorem 2

We prove Theorem 2 generally by showing that the mechanism G designed in Subsection 5.2.1 fully implements the SCF f. In the same manner as the proof of Theorem 1, the proof of Theorem 2 is divided into two parts: "information elicitation" and "implementation with provability."

Part 1 (Information Elicitation): Because H is sufficiently large, we can show in the same manner as in the information elicitation problem (in Subsection 5.1) that any BNE s satisfies

$$I_i^H(s_i^H(\omega_i, t_i)) = \omega_i \text{ for all } i \in N,\ \omega_i \in \Omega_i, \text{ and } t_i \in T_i.$$

Hence, any agent truthfully reveals her private signal at the H-th announcement through $(H_i^H)_{i\in N}$. The iterative elimination method guarantees the existence of strategy profiles that satisfy the Bayesian equilibrium property for the first H sub-messages.

Part 2 (Implementation with Provability): Consider a strategy profile s whose first H announcements satisfy the BNE property. We define the sincere strategy for agent i, which we denote by $\hat{s}_i = (\hat{s}_i^k)_{k=1}^{H+K}$, as

$$\hat{s}^h = s^h \text{ for all } h \in \{1,...,H\},$$

and

$$\hat{s}_i^k(\omega_i, t_i) = \omega_i \text{ for all } k \in \{H+1,...,H+K\}.$$

Clearly, $\hat{s}$ induces the value of the SCF f. Part 2 shows that if s is a BNE, then $s = \hat{s}$ must hold.

Because $I_j^H(m_j^H) = \omega_j$, $p_{i,j}(\cdot\,|\,\omega_i) \neq p_{i,j}(\cdot\,|\,\omega_i')$ for all $\omega_i' \neq \omega_i$, and ξ was selected sufficiently large, it follows from the nature of the quadratic scoring rule $\gamma_{i,j}(m_{i,j}^{H+1}, I_j^H(m_j^H))$ that each agent i is willing to announce $m_{i,j}^{H+1} = \omega_i$ irrespective of the selection of $(m_i^k)_{k=H+2}^{H+K}$. Hence, we have

$$s^{H+1} = \hat{s}^{H+1}.$$

If an agent i announces a sub-message that is different from her $(H+1)-th$ sub-message as the first deviation starting from the $(H+2)-th$ sub-messages, she is fined the monetary amount $\dfrac{\varepsilon}{3+\xi}$. Because we have made K sufficiently large, the impact of the selection of each sub-message on the determination of the allocation is small compared with the monetary amount $\dfrac{\varepsilon}{3+\xi}$. Following Abreu and Matsushima (1992a; 1992b), this drives agents into a tail-chasing competition in a manner similar to the mechanism designed for the proof of Theorem 1, through which each agent avoids becoming the first deviant. Given that we have already proved that all agents announce truthfully at their $(H+1)-th$ sub-messages, this competition drives them to announce the state truthfully from the $(H+2)-th$ sub-message to the $(H+K)-th$ sub-message.

To be precise, consider an arbitrary $k \in \{H+2,...,H+K\}$ and suppose that

$$s^{k'} = \hat{s}^{k'} \text{ for all } k' < k \, .$$

If $m_j^k \neq \omega_j$ for some $j \neq i$, the agent i strictly prefers announcing truthfully at the k-th sub-message, because she can avoid being the first deviant. Even if $m_j^k = \omega_j$ for all $j \neq i$, the agent i still strictly prefers announcing truthfully at the k-th sub-message, because she does not want to increase $r_i(m_i)$ and because the SCF is incentive-compatible. Accordingly, through the iterative elimination of dominated strategies, we can inductively prove that

$$s^k = \hat{s}^k \text{ for all } k \in \{H+2,...,H+K\} \, .$$

Hence, there is no BNE other than $\hat{s}$. Because $\hat{s}$ is a BNE and achieves the value of f, we have completed the proof of Theorem 2.

Remark 5: By assuming that the psychological cost $c_i(m,\omega,t_i,G)$ is convex in $m_i(\omega_i)$, we can replace the full implementation with the unique implementation, which strengthens the results in this section. For the first H sub-messages, an honest agent may have multiple best response sub-messages although any best response guarantees truthful revelation $I_i^H(s_i^H(\omega_i,t_i)) = \omega_i$. However, convexity guarantees the uniqueness of the honest agents' best responses; the SCF is not only fully but also uniquely implementable in the asymmetric information environment.

Remark 6: Incentive compatibility is a necessary condition for implementation, provided that only small fines are permitted. However, if we can utilize large transfers under quasi-linearity, we can prove that any SCF, whether it is incentive compatible or not, is fully implementable. Information diversity implies that

$$p_{i,j}(\cdot \mid \omega_i) \neq p_{i,j}(\cdot \mid \omega_i') \text{ for all } \omega_i' \neq \omega_i ,$$

which guarantees the presence of a payment rule x such that truth-telling is a strict BNE in the information elicitation problem. Hence, we can make any SCF incentive compatible with the help of payment rule design; that is, any SCF is fully implementable.

6. Conclusion

This study investigated a society in which people are either selfish or honest, and showed that every incentive-compatible SCF, whether material or nonmaterial, is implementable in BNE if “all agents are selfish” never happens to be common knowledge.

This study assumed that there exist only two motives for agents: selfishness and honesty. In reality, there could be adversarial motives such as “always tell a lie.” However, our results are robust to the consideration of such motives although it is not explicitly shown in this study. The equilibrium messages are attracted to somewhere close to truth-telling whenever these motives are not as important as honesty; that is, the central planner can still identify the true state by checking whether agents’ messages are attracted by a certain message through the indications $(I_{i,j})$.

Our findings 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?”, “How will decision-making affect outsiders and future generations?”, and others. Selfish people are generally unmotivated by such ethical concerns even if they have a keen interest in ethical concerns and are knowledgeable about them. From the viewpoint of social network (Putnam, 2006) and social common capital (Uzawa, 2005) in epistemology, 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 from each other. With this common knowledge, the central planner cannot derive ethical information from selfish people correctly. However, if selfish people and honest people are path-connected with each other, the central planner can properly derive such information even from selfish agents and reflect it in her normative judgment, that is, she can implement any ethical SCF she desires.

This study considered the role of a social network epistemologically, implicitly assuming that preference for honesty and prosocial propensity are consistent. Consideration of the situation where this premise does not hold is an important issue. The situation where the SCF is given for the central planner’s private purpose, which agents do not agree with, is an example. Matsushima (2013) analyzes the unique implementation in this situation as the possibility of psychological guidance, which causes those agents to reveal information truthfully through a manipulation of the revelation process. Another example is a situation such as scarce resource allocation in a pandemic, where agents have different ethical criteria and the central planner composes an SCF by finding those compromises. The related works are Pathak et al. (2020) and Matsushima (2021b; 2021c). Such research is expected to develop further in the future.

References

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

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. https://www.econexp.org/hitoshi/AMincomplete.pdf

Aoyagi, M. (1998): Correlated Types and Bayesian Incentive Compatible Mechanisms with Budget Balance, Journal of Economic Theory, 79, 142–151.

Arrow, K. (1951): Social Choice and Individual Values, New Haven: 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.

Carlsson, H., and E. van Damme (1993): Global Games and Equilibrium Selection, Econometrica, 61, 989–1018.

Cooke, R. (1991): Experts in Uncertainty: Opinion and Subjective Probability in Science, New York: Oxford University Press.

Charness, G., and M. Dufwenberg (2006): Promises and Partnership, Econometrica, 76 (6), 1579–1601.

Dasgupta, A., and A. Ghosh (2013): Crowdsourced Judgement Elicitation with Endogenous Proficiency, in Proceedings of the 22nd International Conference on World Wide Web, 319–330.

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.

Ellingsen, T., and M. Johannesson (2004): Promises, Threats and Fairness, The Economic Journal, 114 (495), 397–420.

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. Amsterdam: North-Holland.

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

Johnson, S., J. Pratt, and R. Zeckhauser (1990): Efficiency Despite Mutually Payoff-Relevant Private Information: The Finite Case, Econometrica, 58, 873–900.

Kartik, N. (2009): Strategic Communication with Lying Costs, The Review of Economic Studies, 76 (4), 1359–1395.

Kartik, N., M. Ottaviani, and F. Squintani (2007): Credulity, Lies, and Costly Talk, Journal of Economic Theory, 134 (1), 93–116.

Kartik, N., and O. Tercieux (2012): Implementation with Evidence, Theoretical Economics, 7 (2), 323–355.

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

Kong, Y., and G. Schoenebeck (2019): An Information Theoretic Framework for Designing Information Elicitation Mechanisms that Reward Truth-Telling, ACM Transactions on Economics and Computation (TEAC), 7 (1), 1–33.

Kreps, D. M., P. Milgrom, J. Roberts, and R. Wilson (1982): Rational Cooperation in the Finitely Repeated Prisoners' Dilemma, Journal of Economic Theory, 27, 245–252.

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. (1977/1999): 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. (1990): Dominant Strategy Mechanisms with Mutually Payoff-Relevant Information and with Public Information, Economics Letters, 34, 109–112.

Matsushima, H. (1991): Incentive Compatible Mechanisms with Full Transferability, Journal of Economic Theory, 54, 198–203.

Matsushima, H. (1993): Bayesian Monotonicity with Side Payments, Journal of Economic Theory, 59, 107–121.

Matsushima, H. (2007): Mechanism Design with Side Payments: Individual Rationality and Iterative Dominance, Journal of Economic Theory, 133 (1), 1–30.

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

Matsushima, H. (2008b): Role of Honesty in Full Implementation, Journal of Economic Theory, 139, 353–359.

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. (2020): Implementation, Honesty and Common Knowledge, mimeo.

Matsushima, H. (2021a): Partial Ex-Post Verifiability and Unique Implementation of Social Choice Functions, Social Choice and Welfare, 56, 549–567.

Matsushima, H. (2021b): Assignments with Ethical Concerns, Discussion Paper CARF-F-514 (UTMD-007), University of Tokyo.

Matsushima, H. (2021c): Auctions with Ethical Concerns, Discussion Paper CARF-F-515 (UTMD-008), University of Tokyo.

Matsushima, H. and S. Noda (2020): Mechanism Design with Blockchain Enforcement, CARF-F-474, University of Tokyo.

Mazar, N., O. Amir, and D. Ariely (2008): More Ways to Cheat – Expanding the Scope of Dishonesty, Journal of Marketing Research, 45 (6), 651–653.

Miller, N., J. Pratt, R. Zeckhauser, and S. Johnson (2007): Mechanism Design with Multidimensional, Continuous Types and Interdependent Valuations, Journal of Economic Theory, 136 (1), 476–496.

Miller, N., P. Resnick, and R. Zeckhauser (2005): Eliciting Informative Feedback: The Peer-Prediction Method, Management Science, 51 (9), 1359–1373.

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

Morris, S., and H. S. Shin (1998): Unique Equilibrium in a Model of Self-Fulfilling Currency Attacks, American Economic Review, 88 (3), 587–597.

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: Cambridge University Press.

Pathak, P., T. Sonmez, U. Unver, and M. Yenmez (2020): Leaving No Ethical Value Behind: Triage Protocol Design for Pandemic Rationing, SSRN. https://ssrn.com/abstract=3569307

Postlewaite, A., and X. Vives (1987): Bank Runs as an Equilibrium Phenomenon, Journal of Political Economy, 95, 485–491.

Prelec, D. (2004): A Bayesian Truth Serum for Subjective Data, Science, 306 (5695), 462–466.

Putnam, R. (2006): Bowling Alone: Americas's Declining Social Capital, Journal of Democracy, 6(1): 65–78.

Rubinstein, A. (1989): The Electric Mail Game: Strategic Behavior Under ‘Almost Common Knowledge’, *American Economic Review* 79, 385–391.

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.

Uzawa, H. (2005): Economic Analysis of Social Common Capital. Cambridge, England: Cambridge University Press

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