Epistemological Mechanism Design

By

Hitoshi Matsushima (The University of Tokyo)

Shunya Noda (University of British Columbia)

November 2020

CREPE DISCUSSION PAPER NO. 83

CENTER FOR RESEARCH AND EDUCATION FOR POLICY EVALUATION (CREPE)
THE UNIVERSITY OF TOKYO
http://www.crepe.e.u-tokyo.ac.jp/

Epistemological Mechanism Design[1]

Hitoshi Matsushima[2]
University of Tokyo

Shunya Noda[3]
University of British Columbia

First Version: October 26, 2020 This Version: November 4, 2020

Abstract

This study demonstrates a new approach to mechanism design from an epistemological perspective. We introduce an epistemological type space in which agents are either selfish or honest, and show that a slight possibility of honesty in higher-order beliefs motivates all selfish agents to behave sincerely. Specifically, we consider a situation in which a central planner attempts to elicit correct information from informed agents through mutual monitoring. We assume severe restrictions on incentive device availability: neither public monitoring nor allocation rules are available. Thus, the central planner uses only monetary payment rules. If "all agents are selfish" is common knowledge, eliciting correct information as unique equilibrium behavior is generally impossible. However, we show a very permissive result in our epistemological model by designing a quadratic scoring rule as the monetary payment rule: the central planner can elicit correct information from all agents as unique Bayes Nash equilibrium behavior if "all agents are selfish" is never common knowledge. This result holds even if honest agents are mostly motivated by monetary interest.

Keywords: epistemological mechanism design, unique information elicitation, common knowledge of all agents' selfishness, intrinsic preference for honesty, quadratic scoring rule.
JEL Classification Numbers: C72, D71, D78, H41

[1] The first version of this study was entitled "Unique Information Elicitation," CARF-F-496, University of Tokyo, 2020. This is a drastic extension of the theory part of Matsushima and Noda (2020). This study was supported by a grant-in-aid for scientific research (KAKENHI 20H00070) from the Japan Society for the Promotion of Science (JSPS), the Ministry of Education, Culture, Sports, Science and Technology (MEXT) of the Japanese government, and the Social Sciences and Humanities Research Council of Canada.

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

[3] Vancouver School of Economics, University of British Columbia, 6000 Iona Dr, Vancouver, BC V6T 1L4, Canada. E-mail: shunya.noda [at] gmail.com.

1. Introduction

This study demonstrates a new approach to mechanism design from an epistemological perspective. We introduce a type space in which agents are either selfish or honest (nonselfish) in higher-order beliefs. We show that a slight possibility of an honest agent in higher-order beliefs incentivizes all selfish agents to behave sincerely.

We assume that "all agents are selfish," (i.e., all agents are motivated only by their monetary interests) is never common knowledge. Hence, we consider an epistemological possibility that an agent is motivated not only by her (or his) selfish motive but also by ethical or behavioral motives, such as an intrinsic preference for honesty. Most previous studies in the mechanism design literature ignored these motives and assumed that "all agents are selfish" is common knowledge. As a departure from previous literature, this study demonstrates a mechanism design method that helps elicit a nonselfish motive hidden in an agent's mind and harness it to incentivize other selfish agents to behave sincerely.

Specifically, we consider a problem in which a central planner attempts to elicit correct information from agents. The central planner needs to know which state of the world actually occurs, whereas there exists an agent who is fully informed of it. Hence, the central planner attempts to design a mechanism to incentivize this agent to truthfully announce the state. However, we assume severe restrictions on available incentive devices: there are no public monitoring technologies,[4] and the central planner cannot use any allocation device besides monetary transfers.[5] Hence, the central planner in this study is permitted only to use a message-contingent payment rule.

To overcome the difficulty resulting from these restrictions, the central planner listens to the messages from multiple agents who have the same information and have

[4] Several previous works such as the principal-agent problem with hidden information assumed public monitoring technology in which—by observing an ex-post public signal—the central planner confirms, at least in part, whether an agent announces truthfully. See Salanié (1997) for a survey on the principal-agent problem with hidden information.

[5] Several previous works in the literature, such as auction and implementation theory, assumed that multiple tools exist for incentives, such as resource allocation with which the central planner extracts correct information by allowing agents to self-select from multiple menus of allocations and payoffs. See Krishna (2009) for a survey on auction theory and Maskin and Sjöström (2002) and Palfrey (2002) for surveys on implementation theory.

them mutually monitor each other. However, for such mutual monitoring to function, the central planner still needs to overcome another challenge in incentives, that is, the multiplicity of equilibria resulting from coordination failure. Hence, this study articulates the possibility of *unique information elicitation*, implying that the central planner elicits correct information through agents' unique equilibrium behavior.

The mechanism design literature has traditionally assumed that "all agents are selfish" is common knowledge. This assumption makes severe multiplicity of equilibria to be inevitable in our problem because agents' preferences for monetary transfers are independent of the state; therefore, the set of all equilibria is the same across states. However, real people often have nonselfish motives. Hence, the statement derived from this assumption could be useful only if it is robust against the contamination of nonselfish motives.

This study considers the possibility in epistemology that an agent is honest, that is, the agent is motivated by an *intrinsic preference for honesty* as well as monetary interest. Many empirical and experimental studies indicate that human beings are not purely motivated by monetary payoffs but have irresistible preferences for honesty. Abeler et al. (2019) provide a detailed meta-analysis in which they use combined data from 90 studies involving more than 44,000 subjects across 47 countries to show that subjects gave up a large fraction of potential gain from lying.[6]

However, this study allows the case in which an honest agent exists only exceptionally. We allow even honest agents to be motivated by monetary interest; the influences of preferences for honesty on decision making can be arbitrarily small, even if the agent is honest. Importantly, this study does not assume that agents expect the possibility that there exists an honest agent: we allow agents to have mutual knowledge that all agents are selfish.

Despite these weaknesses in honesty, this study shows a very permissive result: *the central planner can overcome this multiplicity and elicit correct information from agents through unique Bayes Nash equilibrium (BNE) behavior if and only if "all agents are selfish" never happens to be common knowledge.* This statement is powerful and

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

profound because it provides a theoretical basis on which a person who commits wrongdoing in the world can be caught only by testimony, ex post facto (limited incentive tools), or without any means of proof (no provability).

In this study, the design of the payment rule has the following characteristics. First, each agent is required to announce not a state but a *distribution of the state*, even if she (or he) is fully informed of the state; she can continuously fine-tune her announcement and payoff. Second, a selfish agent prefers announcing the *same distribution* as the other agents announce in expectation. Third, an honest agent is driven to be *more honest* than a selfish agent. Based on these three characteristics, all agents come to expect the possibility that an agent is driven to be more honest, which drives them into a tail-chasing competition toward honest reporting.

Specifically, we design the payment rule as the *quadratic scoring rule* (Brier, 1950), which aligns agents' payoffs with the distance between their messages. Hence, an agent's monetary payoff is maximized when she reports the average of the other agents' messages. The quadratic scoring rule is one of the standard mechanism design methods in partial implementation with asymmetric information.[7] This study suggests that this method is a powerful solution not only for partial implementation but also for *unique implementation*.

The quadratic scoring rule may put each agent aside various nonselfish motives and prioritize her monetary interest to announce the same messages as those of other agents. However, as Abeler et al. (2019) pointed out, the intrinsic preference for honesty remains unexcluded in this case; therefore, honest agents still have an incentive to announce a little more honestly than selfish agents.

Several studies have investigated the unique (or full) implementation of social choice functions, in which it is assumed that there exists an agent who is honest (Matsushima, 2008a, 2008b; Dutta and Sen, 2012; Kartik et al., 2014). In contrast to these

[7] 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).

works, this study does not assume the existence of honest agents—the only assumption we need is that "all agents are selfish" is not common knowledge.[8]

Just like the email game of Rubinstein (1989), this study contrasts the outcome under common knowledge and "almost common knowledge" of all agents' selfishness (in the sense that honest agents exist only in higher-order beliefs). As we assume that all agents are rational and can correctly distinguish these two, the possibility of honesty in epistemology drives all agents to tell the truth. The email game of Rubinstein (1989) demonstrates that "almost common knowledge" could lead us an unintuitive outcome. While our paper also demonstrates the vulnerability of the common knowledge assumption, the implication of this study is contrasting to Rubinstein's (1989). In an information elicitation mechanism, the intuitive outcome is truthtelling, and people can naturally expect that a truthful strategy profile is a focal point, while there are many equilibria under common knowledge of selfishness. We further show that, by carefully designing a "game" (mechanism), we can eliminate all the unwanted and unintuitive equilibria, whenever "all agents are selfish" is not common knowledge (while it could be "almost common knowledge"). Our result indicates that, when agents believe that others behave honesty "by default" and the mechanism nicely incentivizes agents to coordinate their announcements, then it is difficult for agents to coordinate to jointly deviate to a dishonest reporting.

The equilibrium analysis of the game in the presence of behavioral agents and incomplete information itself has a long history. For example, 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. In contrast, our focus is on the design of mechanisms, which takes full advantage of the possibility of behavioral agents in higher-order beliefs.

This study is organized as follows. Section 2 presents the model, while Section 3 presents the theorem. Section 4 provides an example that outlines the logic behind the

[8] Matsushima (2020b) showed, as an extension of this study, that all social choice functions are uniquely implementable in BNE if "all agents are selfish" is never common knowledge.

theorem and Section 5 provides the proof of the theorem. Section 6 discusses issues such as mixed strategies, budget balancing, number of participants, and robustness against other behavioral motives. Section 7 concludes.

2. The Model

This study investigates a situation in which a central planner attempts to correctly elicit information from multiple agents. Let $\boldsymbol{N} = \{\mathbf{1},\mathbf{2},...,\boldsymbol{n}\}$ denote the finite set of all agents, where $\boldsymbol{n} \geq \mathbf{2}$. Let Ω denote the nonempty and finite sets of possible states. We assume *complete information about the state* across agents. Each agent is informed of the true state $\omega \in \Omega$, whereas the central planner does not know it. Hence, the central planner attempts to design a mechanism that incentivizes these agents to make a truthful announcement.

From an epistemological perspective, we assume that agents are not always *selfish*; that is, they are not always concerned about their monetary interests. Instead, agents could be *honest*, that is, motivated not only by monetary interest but also by an intrinsic preference for honesty. To be precise, we assume *incomplete information concerning honesty* in that each agent knows whether they are selfish or honest, but the other agents are not informed of it. To formulate this incomplete information, we define the *type space* as follows, which is based on Bergemann and Morris (2005, 2012):

$$\Gamma \equiv (\boldsymbol{T}_i,\pi_i,\theta_i)_{i\in N},$$

where $\boldsymbol{t}_i \in \boldsymbol{T}_i$ is agent $\boldsymbol{i}$'s type, $\theta_i : \Omega \times \boldsymbol{T}_i \to \{\mathbf{0},\mathbf{1}\}$, and $\pi_i : \Omega \times \boldsymbol{T}_i \to \Delta(\boldsymbol{T}_{-i})$.[9] Each agent $\boldsymbol{i}$ knows her type $\boldsymbol{t}_i$ and the state ω, but does not know the other agents' types $\boldsymbol{t}_{-i}$. Agent $\boldsymbol{i}$ expects that the other agents' types are distributed according to a probability measure $\pi_i(\omega,\boldsymbol{t}_i) \in \Delta(\boldsymbol{T}_{-i})$. Each agent is either selfish or honest: agent $\boldsymbol{i}$ is selfish (honest) if $\theta_i(\omega,\boldsymbol{t}_i) = \mathbf{0}$ ($\theta_i(\omega,\boldsymbol{t}_i) = \mathbf{1}$, respectively). More details will be subsequently explained.

[9] We denote by $\Delta(\boldsymbol{Z})$ the space of probability measures on the Borel field of a measurable space $\boldsymbol{Z}$. We denote $\boldsymbol{Z} \equiv \underset{i\in N}{\times} \boldsymbol{Z}_i$, $\boldsymbol{Z}_{-i} \equiv \underset{j\neq i}{\times} \boldsymbol{Z}_j$, $z = (z_i)_{i\in N} \in \boldsymbol{Z}$, and $z_{-i} = (z_j)_{j\neq i} \in \boldsymbol{Z}_{-i}$.

The central planner designs a mechanism $G \equiv (M, x)$, where $M = \underset{i \in N}{\times} M_i$ $x = (x_i)_{i \in N}$ denotes a *payment rule* and $x_i : M \to R$ denotes the payment rule for agent i. Each agent i simultaneously announces a message $m_i \in M_i$ and obtains a monetary payment $x_i(m) \in R$ from the central planner, where we denote $m = (m_j)_{j \in N} \in M$.

We consider a class of indirect mechanisms in which each agent announces a *probability distribution over states* as the message, that is,

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

We write $m_i = \omega$ if $m_i(\omega) = 1$. 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 the probability distribution over states $m_i = s_i(\omega, t_i) \in \Delta(\Omega)$ when the state $\omega \in \Omega$ occurs.

If agent i is selfish, they maximize the expected value of their utility given by monetary payment $x_i(m)$:

$$[\theta_i(\omega, t_i) = 0] \Rightarrow [\text{agent } i \text{ selects}$$

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

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

In contrast, if agent i is *honest*, they are motivated not only by monetary interest but also by an intrinsic preference for honesty and maximizes *the expected payment minus their psychological cost*:

$$[\theta_i(\omega, t_i) = 1] \Rightarrow [\text{agent } i \text{ selects}$$

$$m_i = s_i(\omega, t_i) \in \arg\max_{m_i \in M_i} E[x_i(m_i, s_{-i}(\omega, t_{-i}))$$

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

where $c_i(m, \omega, t_i, G) \in R$ denotes their psychological cost. We assume that c_i represents the intrinsic preference for honesty. Specifically, for every $i \in N$, $\omega \in \Omega$, $m \in M$, and $\tilde{m}_i \in M_i$,

(1) $$[\theta_i(\omega, t_i) = 1,\ m_i(\omega) > m_i'(\omega), \text{ and } x_i(\tilde{m}_i, m_{-i}) > x_i(m)]$$

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

Assumption (1) implies that any honest agent feels more or less guilty about telling lies that generate more self-interest. Hence, any honest agent strictly prefers making an announcement more honestly than the selfish types. In this study, we allow each agent's psychological cost to be arbitrarily small, even if this agent is honest: we do not set any condition on how much an agent cares about honesty.

An example of psychological cost is given by

$$c_i(m,\omega,t_i,G) = \lambda_i\{1 - m_i(\omega)\},$$

where $\lambda_i > 0$. This example describes the preference for honesty with which an agent can save psychological cost by making an announcement more honestly.

Another example is given by

$$c_i(m,\omega,t_i,G) = \max[0, x_i(m) - x_i(\tilde{m}_i,m_{-i})]\lambda_i\{1 - m_i(\omega)\},$$

where $\tilde{m}_i = \omega$ and $\lambda_i > 0$. In this example, the psychological cost depends crucially on the shape of the payment rule: the magnitude of the impact of the lie on their monetary payoff influences the size of the psychological cost.

In both examples, by setting λ_i close to zero, we can consider the case in which the direct impact of the preference for honesty on an agent's decision-making can be arbitrarily small. In such a case, even honest agents are mostly motivated by monetary interests. As implied by the latter example, we can also consider the case in which the direct impact of preference for honesty on an agent's decision-making is arbitrarily small compared with the impact of the lie on their monetary payoff.

This study investigates BNE in a game associated with a payment rule x.

3. The Theorem

We specify the payment rule $x = x^*$ as the following *quadratic scoring rule*: for every $i \in N$ and $m \in M$,

$$x_i^*(m) = -\sum_{j\neq i}[\sum_{\omega\in\Omega}\{m_i(\omega) - m_j(\omega)\}^2],$$

which describes the distance of agent $i's$ message from the other agents' messages. From simple calculations, if s is a BNE in the game associated with x^*, then for every $i \in N$ and $(\omega, t_i) \in \Omega \times T_i$,

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

whereas

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

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

That is, *any selfish agent mimics the average of the other agents' announcements in expectation, whereas any honest agent makes announcements more honestly than a selfish agent.*

We define the *truthful strategy* profile s^* by

$$s_i^*(\omega,t_i) = \omega \quad \text{for all } \; i \in N \; \text{ and } \; (\omega,t_i) \in \Omega \times T_i,$$

according to which each agent i announces truthfully about the state irrespective of the state and type. We consider a necessary and sufficient condition under which the truthful strategy profile s^* is the unique BNE in the game associated with x^*; that is, the central planner succeeds in eliciting correct information about the state from the agents as unique equilibrium behavior.

We call a subset of type profiles $T \equiv \underset{i\in N}{\times} T_i$ an *event*. For convenience, for each event $E \subset T$, we write

$$\pi_i(E \mid \omega,t_i) = \pi_i(E_{-i}(t_i) \mid \omega, t_i),$$

where we denoted $E_{-i}(t_i) \equiv \{t_{-i} \in T_{-i} \mid (t_i, t_{-i}) \in E\}$. 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

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

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 then define

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

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

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

Note that if E is common knowledge at (ω,t), then

$$\pi_i(\underset{j\in N}{\times} V_j^\infty(E,\omega) \mid \omega,t_i) = 1 \text{ for all } i \in N.$$

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: *The truthful strategy profile* s^* *is the unique BNE in the game associated with* x^* *if and only if*

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

From the definition of common knowledge, the necessary and sufficient condition of the theorem clearly implies that $V_i^\infty(E^*(\omega),\omega) = \phi$ for all $i \in N$ and $\omega \in \Omega$. The theorem states that *all agents, whether selfish or honest, are willing to announce the state truthfully as unique BNE behavior if and only if "all agents are selfish" is never common knowledge.* Hence, with the elimination of common knowledge of all agents' selfishness, the central planner can always succeed in eliciting correct information about the state from agents. We should regard this elimination as the *minimal* requirement of an epistemological potential that an agent cares about honesty. In fact, the success of correct elicitation holds even if "all agents are selfish" is mutual knowledge.

4. Example

The following characteristics of the quadratic scoring rule x^* are crucial for understanding the theorem. For simplicity of the arguments, we focus on the two-agent case.

(a) Each agent's message space is not the set of states but the set of probability distributions over states. Hence, an agent can continuously fine-tune their message and payment.

(b) Any selfish agent is incentivized to report the same distribution as the other agent's report in expectation.

(c) Suppose that agent 1 is selfish and expects the possibility that agent 2 makes an announcement more honestly than what agent 2 expects about the announcement by agent 1 with selfish type. Then, agent 1 with selfish type has an incentive to make the announcement (slightly) more honestly than agent 2 expects. The same scenario holds true even if agents 1 and 2 are replaced. This will be the driving force for a tail-chasing competition through which each agent announces more honestly than the other, reaching honest reporting by both.

Trivially, whenever agent i expects the possibility that the other agent $j \neq i$ is honest, then the supposition in (c) holds and agent i is driven to be more honest. However, the other agent j does not have to be honest: it is necessary and sufficient that agent i expects the possibility that the other agent j, whether selfish or honest, is driven to be more honest.

Let us propose the following example with a finite type space, where $n = 2$, $\Omega = \{0,1\}$, and $T_i = \{0,1,...,H\}$ for each $i \in \{1,2\}$. We assume that agent i is honest if and only if $t_i = 0$, that is,

$$[\theta_i(\omega, t_i) = 0] \Leftrightarrow [t_i = 0].$$

The message space of agent i is given by $M_i = [0,1]$, where $m_i \in [0,1]$ indicates the probability that state 1 ($\omega = 1$) occurs. The quadratic scoring rule is given by $x_1^*(m) = x_2^*(m) = -(m_1 - m_2)^2$.

We assume that there exists a common prior over-type profile, π , and it is symmetric; that is, $\pi(h,h') = \pi(h',h)$ for all $(h,h') \in \{0,1,...,H\}^2$. Because the mechanism and agents are symmetric, we often refer to an agent with type h as a "type-h agent" without specifying their identity $i \in \{1,2\}$. We assume that the set of selfish types $\{1,\ldots,H\}$ is (weakly) connected in the sense that

$$\pi(h,h+1) > 0 \text{ for all } h \in \{1,...,H-1\}.$$

Without loss of generality, we assume that the true state is $\omega = 1$ (the analysis for the case of $\omega = 0$ is similar), and we drop it from the notation. The psychological cost for each agent i with honest type is given by $\lambda(1-m_i)$, where $\lambda > 0$. Let $\bar{m}_j(t_i;s_j)$ be agent $j's$ expected message conditional on agent $i's$ type t_i:

$$\bar{m}_j(t_i;s_j) \equiv E\left[s_j(t_j) \mid t_i\right] = \sum_{h=0}^{H} \pi_i(h \mid t_i) s_j(h).$$

Then, agent $i's$ best response against s_j is given by

$$BR_i(s_{-i},t_i) = \begin{cases} \bar{m}_j(t_i;s_j) & \text{if } t_i \in \{1,\ldots,H\} \\ \min\left\{\bar{m}_j(t_i;s_j) + \lambda/2, 1\right\} & \text{if } t_i = 0 \end{cases}.$$

Hence, any honest agent is driven to be more honest than a selfish agent.

Case 1: First, consider the case in which the set of selfish types is disconnected from the honest type, that is,

$$\pi(0,h) = 0 \text{ for all } h \in \{1,...,H\}.$$

Any selfish agent expects that the other agent is selfish with certainty, and any honest agent expects that the other agent is honest with certainty.

When $t = (0,0)$ is realized, the best response of each agent $i \in \{1,2\}$ is given by $s_i(0) = \min\{s_j(0) + \lambda/2, 1\}$; that is, the preference for honesty drives each agent to select a message that is slightly more honest than the other. Clearly, in a BNE s, $s_1(0) = s_2(0) = 1$ must be satisfied.

In contrast, an equilibrium strategy can take any value when $h \in \{1,...,H\}$. As long as there exists a constant $p \in [0,1]$ such that

$$s_i(h) = p \text{ for all } i \in N \text{ and } h \in \{1,\dots,H\},$$

it is a BNE. Hence, there are infinitely many BNEs in which any selfish agent may tell a lie. Clearly, we fail to elicit the correct state as a unique BNE in Case 1.

Case 2: Consider the case in which, unlike Case 1, the set of selfish types is connected with the honest type in a minimal sense such that there exists $h \in \{1,...,H\}$ with $\pi(0,h) > 0$. For simplicity, we assume that $h = 1$, that is,

$$\pi(0,1) > 0.$$

It is easy to see that the same argument holds true even if we replace type 1 with any $h \in \{2,...,H\}$.

Because of higher-order reasoning, this minimal connection drastically changes the set of BNEs as follows. Clearly, a type-0 (honest) agent is driven to be more honest. The minimal connection implies that a type-1 agent expects that the other agent may be type-0 with a positive probability, and she would like to match her message with the other agent's announcement in expectation. Hence, the type-1 agent is also driven to be more honest. Similarly, a type-2 agent expects that the other agent may be type-1 with a positive probability and, thus, is driven to be more honest. We can iterate this argument and verify that any agent, whether selfish or honest, is driven to be more honest, that is, attempts to send a more honest message than the other. This structure of best responses immediately leads us to the uniqueness of BNE, where all agents report truthfully.

Note that this uniqueness holds even if both agents' selfishness is mutual knowledge. As long as $t_1 \geq 2$ and $t_2 \geq 2$, each agent does not expect that the other agent may be honest. However, the aforementioned higher-order reasoning will guide any agent to send a more truthful message, which drastically shrinks the set of BNE. As long as there is no common knowledge of both agents' selfishness, this logic always functions and the uniqueness of the BNE is guaranteed.

Case 1 corresponds to situations in which all selfish types completely eliminate associations with honest types. In this case, unique information elicitation is impossible. However, if there is at least one selfish type who expects even a little (possibly indirect) influence of an honest type, then unique information elicitation is achievable. The driving force behind this phenomenon is not that more people become honest but that *selfish people do not rule out the existence of honest agents from their epistemological considerations.*

5. Proof of the Theorem

It is clear from (2) and (3) that s^* is a BNE. Suppose that s is a BNE. Fix an arbitrary $\omega \in \Omega$. Let

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

and

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

Suppose that

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

From the definition of common knowledge, this supposition implies that

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

Toward a contradiction, suppose that

$$\alpha < 1,$$

which implies that there exists a type that announces dishonestly. Note from (2) and (3) that any honest agent prefers making announcements more honestly than selfish agents, implying that no honest type belongs to $\tilde{T}_i$:

$$[t_i \in \tilde{T}_i] \Rightarrow [\theta_i(\omega, t_i) = 0].$$

Consider an arbitrary $i \in N$ and $t_i \in \tilde{T}_i$. From (2) and (3), α equals the average of the other agents' announcements on ω in expectation but not greater than any announcement. Hence, type t_i expects that any other agent $j \neq i$ announces $m_j(\omega) = \alpha$, that is,

$$\pi_i(\underset{j\in N}{\times}\tilde{T}_j \mid \omega,t_i)=1 .$$

This, along with (2) and (3), implies that agent i with type t_i expects that the other agents are never honest, that is,

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

Hence, we have

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

Moreover, because

$$\pi_i(\underset{j\in N}{\times} V_i^1(E^*(\omega),\omega)\mid \omega,t_i)\geq \pi_i(\underset{j\in N}{\times}\tilde{T}_i \mid \omega,t_i)=1 ,$$

we have $\pi_i(\underset{j\in N}{\times} V_i^1(E^*(\omega),\omega)\mid \omega,t_i)=1$, that is,

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

Similarly, we have

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

Hence, we have

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

which however contradicts the supposition that $V_i^\infty(E^*(\omega),\omega)=\phi$. Hence, we conclude that $\alpha=1$ or, equivalently, $s=s^*$, must be the case in any BNE; thus, we have proved the "if" part of the theorem.

Fix an arbitrary $\omega'\neq\omega$. We specify a strategy profile s^+ as follows: for every $i\in N$ and $t_i\in T_i$,

$$s_i^+(\omega,t_i)=\omega \qquad \text{if } t_i\notin V_i^\infty(E^*(\omega),\omega) ,$$

$$s_i^+(\omega,t_i)=\omega' \qquad \text{if } t_i\in V_i^\infty(E^*(\omega),\omega) ,$$

and

$$s_i^+(\tilde{\omega},t_i)=s_i^*(\tilde{\omega},t_i) \text{ for all } \tilde{\omega}\neq\omega .$$

It is clear from (2) and the previous argument that s^+ is a BNE, and $s^+\neq s^*$ whenever $V_i^\infty(E^*(\omega),\omega)\neq\phi$ for some $i\in N$. Hence, we have proven the "only-if" part of the theorem.

6. Discussion

6.1. Mixed Strategies

This study considered only pure strategy BNE. However, we can directly use the same logic for the uniqueness of the *mixed strategy* BNE. Because of the quadratic scoring rule, irrespective of whether the other agents' strategies are mixed or pure, any selfish agent prefers announcing the same distribution as the other agents' announcements in expectation, whereas any honest agent prefers announcing more honestly than a selfish agent. The resultant tail-chasing competition eliminates any unwanted BNE, whether pure or mixed.

6.2. Budget-Balancing

The quadratic scoring rule $\boldsymbol{x}^*$ does not balance the budget. In fact, in the two-agent case, it is difficult to find an alternative rule that induces unique information elicitation $\boldsymbol{x}^*$ and balances the budget. In contrast, in the three-or-more-agent case, it is easy to check that the following payment rule $\boldsymbol{x}^+$ induces unique information elicitation and satisfies the budget-balance property: for every $\boldsymbol{i} \in \boldsymbol{N}$ and $\boldsymbol{m} \in \boldsymbol{M}$,

$$\boldsymbol{x}_i^+(\boldsymbol{m}) = \boldsymbol{x}_i^*(\boldsymbol{m}) + \boldsymbol{r}_i(\boldsymbol{m}_{-i}),$$

where

$$\boldsymbol{r}_i(\boldsymbol{m}_{-i}) \equiv \frac{\mathbf{1}}{\boldsymbol{n}-\mathbf{2}} \sum_{i' \neq i, j \neq i, i' \neq j} [\sum_{\omega \in \Omega} \{\boldsymbol{m}_{i'}(\omega) - \boldsymbol{m}_j(\omega)\}^2].$$

6.3. Number of Participants

The theorem holds irrespective of the number of agents participating in the central planner's problem. However, informally, the restrictiveness of the necessary and sufficient conditions depends on this number. In other words, the more agents participate in the problem, the less likely it is that "all agents are selfish" is common knowledge.

If the number of participants is limited, the central planner should recruit informed people from a wider range. If an agent is selfish and believes that other agents are similar, the agent is likely to expect common knowledge of all agents' selfishness. In such a case, the set of selfish types becomes disconnected from the honest type and unique information elicitation may fail. If participants have diverse backgrounds, then selfish agents may expect that the others have different preferences and beliefs and truthful messages could be induced. Let us return to the example in Section 4. Suppose that the central planner picks up agent 1 from a narrower range than what Case 2 assumes, eliminating the possibility that type 0 is chosen. Then, the set of selfish types becomes disconnected from the honest type; thus, unique information elicitation is impossible.

6.4. Other Behavioral Motives

This study assumed that there exist only two categories of agents: selfish agents and honest agents. However, in reality, there are potentially various irrational motives, such as "always tell a lie," "always announce 1," and "always announce 0" in the example of Section 4. If we explicitly consider these motives, we can no longer show that a selfish agent is attracted to announce a literally truthful message. Nevertheless, the equilibrium messages are attracted to somewhere close to truth-telling whenever these motives are not as important as honesty. In the example, because of the extension of the message space from the binary set $\{0,1\}$ to the interval $[0,1]$, the central planner can identify the true state by checking whether agents' messages are closer to 0 or 1.

Abeler et al. (2019) empirically and experimentally suggested that intrinsic preferences for honesty are the main motivation in a wide range of observed behaviors. Their result supports the assumption that each agent is either selfish or honest. However, the motive that is dominant for an agent may depend on the context, and the central planner had better select agents from a population that has little to do with her purpose.

We also point out that agents' behavioral motivations could be influenced by the scale of the payment rules. More specifically, when a central planner sets up an excessively small-scale payment rule, agents may disrespect the mechanism and may not be motivated to behave honestly (such a psychological cost function is considered in the second example of Section 2). In such a case, the intrinsic preference for honesty could be weakened, and other behavioral motivates could significantly affect agents' behavior. Recall that the theorem itself assumes no condition on the scale of the quadratic scoring rule. Accordingly, if the central planner can tune the scale to enhance the preference for honesty, then the payment scale could be selected in such a way that honesty dominates the other behavioral motives.

6.5. Blockchain and Oracle Problem

This study assumed that the central planner has the power to force payments according to the predetermined mechanism (quadratic scoring rule). However, a companion work (Matsushima and Noda, 2020) points out that the argument in this study does not depend on the presence of such a central planner or the court; without external coercion, we can automate and self-enforce the monetary payment rule within the scope of current digital technology. That is, by using digital currencies, the message-contingent monetary payment rule can be computer-programmed as a so-called *smart contract* and deployed on a blockchain such as Ethereum. However, in this case, we face the problem of how to incentivize agents to input correct information into the smart contract—the *oracle problem* in the blockchain literature. This problem is regarded as one of the most important problems that hinders the effective use of smart contracts. Matsushima and Noda (2020) show that this study's theorem provides a new and promising direction to solve this problem.

7. Concluding Remarks

We investigated unique information elicitation, in which the central planner uses only payment rules and agents are either selfish or honest. We proved that the central planner could elicit correct information from agents through their unique BNE behavior if and only if "all agents are selfish" is never common knowledge.

Investigating asymmetric information, where agents can access their respective private information channels, is an important extension. Can quadratic scoring rules still function? If not, what is the design of the payment rule that solves unique information elicitation? How do we define the intrinsic preference for honesty in this environment? Is the exclusion of common knowledge of all agents' selfishness still sufficient for the elicitation of unique information? In this research, these questions represent only the tip of the iceberg but could include new theoretical substances beyond the scope of this study.[10]

References

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

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

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

[10] A special case of asymmetric information is the environment in which each agent fails to access the full information channel with a positive probability. Appendix A in the supplement provides an additional argument and shows that the theorem can be directly applied to this case provided uninformed agents are never selfish. Appendix B in the supplement provides an example that expresses some difficulty in asymmetric information environments.

World Wide Web, 319–330.

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.

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. Shoenebeck (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.

Koszegi, B. (2014): Behavioral Contract Theory, Journal of Economic Literature, 52 (4), 1075–1118.

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.

Krishna, V. (2009): Auction Theory, Academic Press.

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): Role of Honesty in Full Implementation, Journal of Economic Theory, 139, 353–359.

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

Matsushima, H. (2020a): Implementation without Expected Utility: Ex-Post Verifiability, Social Choice and Welfare, 53 (4), 575–585.

Matsushima, H. (2020b): Implementation, Honesty and Common Knowledge, mimeograph.

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.

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

Palfrey, T. R. (2002): Implementation theory, Handbook of Game Theory with Economic Applications 3, 2271–2326.

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.

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

Salanié, B. (1997): The Economics of Contracts: A Primer, Cambridge, MA: MIT Press.

Supplement: Appendices A and B

Appendix A: Uncertainty in Information Access

Appendix A considers the case in which the central planner does not know if each agent can access the information channel. We modify the type space as follows:

$$\Gamma \equiv (T_i, \pi_i, \theta_i, \eta_i)_{i \in N}, \text{ where } \eta_i : T_i \to \{0,1\}.$$

The agent i is informed (uninformed) if $\eta_i(t_i) = 0$ ($\eta_i(t_i) = 1$, respectively). Agent i with $\eta_i(t_i) = 0$ ($\eta_i(t_i) = 1$) can (cannot, respectively) access the information channel and observe the state. We assume that any agent always expects that other agents exist and are informed with a positive probability.

Assumption A-1: For every $i \in N$ and $(\omega, t_i) \in \Omega \times T_i$,

$$\pi_i(\{t_{-i} \in T_{-i} \mid \theta_j(\omega, t_j) = 0 \text{ for some } j \neq i\} \mid \omega, t_i) > 0.$$

We also assume that any uninformed agent is honest. Hence, we categorize agents into three cases, that is, “selfish and informed,” “honest and informed,” and “honest and uninformed.”

Assumption A-2: For every $i \in N$ and $(\omega, t_i) \in \Omega \times T_i$,

$$[\pi_i(t_i) = 1], \Rightarrow [\theta_i(\omega, t_i) = 0].$$

We consider a class of mechanisms (M, x) where each agent announces either a probability distribution on Ω or “μ”:

$$M_i = \Delta(\Omega) \cup \{\mu\} \text{ for all } i \in N.$$

The announcement of “μ” implies that she is uninformed.

If agent i is selfish and informed, they maximize the expected payment:

$[\theta_i(\omega, t_i) = 0$ and $\eta_i(t_i) = 0]$

$\Rightarrow$ [agent i selects

$$m_i = s_i(\omega,t_i) \in \arg\max_{m_i \in M_i} E[x_i(m_i, s_{-i}(\omega,t_{-i})) \mid \omega, t_i]].$$

If agent i is honest and informed, they maximize the expected payment minus the psychological cost:

$$[\theta_i(\omega,t_i) = 1 \text{ and } \eta_i(t_i) = 0]$$

$$\Rightarrow \text{ [agent } i \text{ selects } m_i = s_i(\omega,t_i)$$

$$\in \arg\max_{m_i \in M_i} E[x_i(m_i, s_{-i}(\omega,t_{-i})) - c_i(m_i,\omega,t_i) \mid \omega, t_i]].$$

We assume that

$$[\theta_i(\omega,t_i) = 1,\ \eta_i(t_i) = 0, \text{ and } m_i(\omega) > m_i'(\omega)]$$

$$\Rightarrow [c_i(m_i,\omega,t_i) < c_i(m_i',\omega,t_i)],$$

and

$$[\theta_i(\omega,t_i) = 1 \text{ and } \eta_i(t_i) = 0]$$

$$\Rightarrow [c_i(\mu,\omega,t_i) \geq c_i(m_i,\omega,t_i) \text{ for all } m_i \neq \mu].$$

Given the latter inequalities, pretending to be uninformed is more dishonest than incorrectly announcing about the state. For simplicity of arguments, we assume that if agent i is honest and uninformed, they announce μ :

$$[\theta_i(\omega,t_i) = 1 \text{ and } \eta_i(t_i) = 1]$$

$$\Rightarrow \text{ [The agent } i \text{ selects } m_i = s_i(\omega,t_i) = \mu].$$

We specify the payment rule x^{**} as a modification of x^*. For every $i \in N$ and $m \in M$,

$$x_i^{**}(m) = -\sum_{\substack{j \neq i \\ m_j \neq \mu}} [\sum_{\omega \in \Omega} \{m_i(\omega) - m_j(\omega)\}^2]$$

$$\text{if } m_i \in \Delta(\Omega),$$

and

$$x_i^{**}(m) = -\varepsilon - \max_{\tilde{m}_i \in \Delta(\Omega)} \sum_{\substack{j \neq i \\ m_j \in \Delta(\Omega)}} [\sum_{\omega \in \Omega} \{\tilde{m}_i(\omega) - m_j(\omega)\}^2]$$

$$\text{if } m_i = \mu,$$

where $\varepsilon > 0$. Note that

$$\min_{m_i \in \Delta(\Omega)} x_i^{**}(m) = x_i^{**}(\mu, m_{-i}) + \varepsilon \quad \text{for all } m_{-i} \in M_{-i}.$$

From these specifications and assumptions, if s is a BNE in the game associated with x^{**}, then for every $i \in N$ and $(\omega, t_i) \in \Omega \times T_i$,

(A-1) $$[\theta_i(\omega, t_i) = 0 \text{ and } \eta_i(t_i) = 0]$$

$$\Rightarrow \quad [s_i(\omega, t_i) = E[\frac{\sum_{j \neq i, m_j \neq \mu} s_j(\omega, t_j)}{|\{j \neq i \mid m_j \neq \mu\}|} \mid \omega, t_i]],$$

(A-2) $$[\theta_i(\omega, t_i) = 1 \text{ and } \eta_i(t_i) = 0]$$

$$\Rightarrow \quad [\text{either } s_i(\omega, t_i)(\omega) = 1 \text{ or}$$

$$s_i(\omega, t_i)(\omega) > E[\frac{\sum_{j \neq i, m_j \neq \mu} s_j(\omega, t_j)(\omega)}{|\{j \neq i \mid m_j \neq \mu\}|} \mid \omega, t_i]],$$

and

$$[\theta_i(\omega, t_i) = 1 \text{ and } \eta_i(t_i) = 1]$$

$$\Rightarrow \quad [s_i(\omega, t_i) = \mu].$$

Any selfish and informed agent mimics the average of the other informed agents' announcements in expectation. Any honest and informed agent announces facts more honestly than selfish agents. Any honest and uninformed agent truthfully reports the fact that they are uninformed.

We define the truthful strategy profile s^{**} as follows: for every $i \in N$ and $(\omega, t_i) \in \Omega \times T_i$,

$$s_i^{**}(\omega, t_i)(\omega) = 1 \qquad \text{if } \eta_i(t_i) = 0,$$

and

$$s_i^{**}(\omega, t_i) = \mu \qquad \text{if } \eta_i(t_i) = 1.$$

We denote by $E^{**}(\omega) \subset T$ the event that the state ω occurs and there exists no agent who is honest and informed, that is,

$$E^{**}(\omega) \equiv \{t \in T \mid \forall i \in N : (\theta_i(\omega, t_i), \eta_i(t_i)) \neq (1, 0)\}.$$

Theorem A: *The truthful strategy profile* s^{**} *is the unique BNE in the game associated with* x^{**} *if and only if*

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

Proof: It is clear that s^{**} is a BNE. Suppose that s is a BNE. Fix an arbitrary $\omega\in\Omega$. Let

$$\alpha\equiv\min_{(i,t_i),\eta_i(t_i)=0} s_i(\omega,t_i)(\omega),$$

and

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

Suppose that $V_i^\infty(E^{**}(\omega),\omega)=\phi$ for all $i\in N$, and

$$\alpha<1,$$

which implies that there exists a type who is informed and announces facts dishonestly. Note

$$[t_i\in\tilde{T}_i],\ \Rightarrow[\theta_i(\omega,t_i)=0].$$

Consider an arbitrary $i\in N$ and $t_i\in\tilde{T}_i$. From (A-1) and (A-2), α equals the average of the other agents' announcements on ω in expectation but not greater than any announcement. Hence, agent i expects that any other informed agent $j\neq i$ announces $m_j(\omega)=\alpha$, that is,

$$\pi_i(\underset{j\in N}{\times}(\tilde{T}_j\cup\hat{T}_j)\,|\,\omega,t_i)=1,$$

where $\hat{T}_j$ denotes the set of agent $j's$ types that are uninformed, that is,

$$\hat{T}_j\equiv\{t_j\in T_j\,|\,\eta_j(t_j)=1\}.$$

Because $\underset{j\in N}{\times}(\tilde{T}_j\cup\hat{T}_j)\subset E^{**}(\omega)$, agent i expects the other agents to be either selfish or uninformed, that is,

$$\pi_i(E^{**}(\omega)\,|\,\omega,t_i)=1.$$

Hence, we have

$$\underset{j\in N}{\times}(\tilde{T}_j\cup\hat{T}_j)\subset V_i^1(E^{**}(\omega),\omega).$$

Moreover, because

$$\pi_i(\underset{j\in N}{\times} V_i^1(E^{**}(\omega),\omega) \mid \omega, t_i) \geq \pi_i(\underset{j\in N}{\times}(\tilde{T}_j \cup \hat{T}_j) \mid \omega, t_i) = 1,$$

we have $\pi_i(\underset{j\in N}{\times} V_i^1(E^{**}(\omega),\omega) \mid \omega, t_i) = 1$, that is,

$$\underset{j\in N}{\times}(\tilde{T}_j \cup \hat{T}_j) \subset V_i^2(E^{**}(\omega),\omega).$$

Similarly, we have

$$\underset{j\in N}{\times}(\tilde{T}_j \cup \hat{T}_j) \subset V_i^k(E^{**}(\omega),\omega) \text{ for all } k \geq 2.$$

Hence, we have

$$\underset{j\in N}{\times}(\tilde{T}_j \cup \hat{T}_j) \subset V_i^\infty(E^{**}(\omega),\omega),$$

which, however, contradicts the supposition that $V_i^\infty(E^{**}(\omega),\omega) = \phi$. Hence, we conclude $\alpha = 1$, that is, $s = s^{**}$, and thus, we have proved the "if" part.

Q.E.D.

Appendix B: Difficulty in Asymmetric Information

Let us consider an example with finite type space: $n = 2$, $\Omega = \Omega_1 \times \Omega_2 = \{0,1\}^2$, and $T_i = \{0,1\}$, where each agent $i \in \{0,1\}$ privately observes $\omega_i \in \{0,1\}$ with equal probability, and $t_i = 0$ ($t_i = 1$) implies that agent i is selfish (honest). Each agent i announces a message $m_i \in [0,1]$, implying a probability of $\omega_i = 0$. Consider the following payment rule:

$$x_1^*(m) = x_2^*(m) = -(m_1 - m_2)^2.$$

For simplicity, we assume that any honest type of agent i announces $m_i = \omega_i$, whereas any selfish type maximizes the expected payment. Each agent $i's$ expectation about the other agent $j's$ private information and type is described by $\pi_i(\omega_j, t_j \mid \omega_i, t_i)$, which implies the probability that the other agent j observes private information ω_j and has type t_j, provided that payer i observes private information ω_i and has type t_i.

Assume that there exist $\boldsymbol{q^0 > 0}$ and $\boldsymbol{q^1 > 0}$ such that

$$\boldsymbol{q^0 = \pi_1(0,1\,|\,0,0) = \pi_1(0,1\,|\,1,0) = \pi_2(0,1\,|\,0,0) = \pi_2(0,1\,|\,1,0)},$$

and

$$\boldsymbol{q^1 = \pi_1(1,1\,|\,0,0) = \pi_1(1,1\,|\,1,0) = \pi_2(1,1\,|\,0,0) = \pi_2(1,1\,|\,1,0)}.$$

This assumption implies that whenever each agent’s private information is perfectly correlated, that is, in the complete information environment of the state, then it eliminates the common knowledge of all agents’ selfishness.

Let

$$\boldsymbol{p = \frac{q^1}{q^0 + q^1}}.$$

We can show that any selfish agent $\boldsymbol{i's}$ announcing $\boldsymbol{m_i = p}$ regardless of the realization of ω_i is a BNE. Hence, the central planner fails to elicit correct information from selfish agents, even if the common knowledge of all agents’ selfishness is eliminated.