Epistemological Mechanism Design (Extended Abstract)*

Hitoshi Matsushima[1] and Shunya Noda[2]

[1] University of Tokyo, Japan hitoshi@e.u-tokyo.ac.jp..
[2] University of British Columbia, Canada shunya.noda@gmail.com.

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 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 impossible. Nevertheless, we show a very permissive result: 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 interests. Our method can be applied to solve the oracle problem of the smart contract design.

Keywords: unique information elicitation · common knowledge of all agents' selfishness · intrinsic preference for honesty · quadratic scoring rule · oracle problem

1 Introduction

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, and the central planner cannot use any allocation device besides monetary transfers. Hence, the central planner in this study is permitted only to use a message-contingent payment rule.

* 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.

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 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 intrinsic preferences for honesty. [1] 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.

However, this study allows the case in which an honest agent exists only exceptionally. We allow honest agents to be motivated mostly by monetary interest; the influences of preferences for honesty on decision making can be arbitrarily small, even if the agent is honest. Furthermore, 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 (i.e., all agents know that all agents are selfish).

Despite these weaknesses in honesty, this study shows a very permissive result: the central planner can overcome the multiplicity of equilibria 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 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*, while she (or he) is fully informed of the state; she can continuously fine-tune her announcement and payoff. Second, a selfish agent is incentivized to match her message with the other agents. Third, an honest agent is driven to be *more honest* than a selfish agent. Due to these three characteristics, all agents come to expect the possibility that some of other agents are 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* [4], 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.[3] This study suggests that this method is a powerful solution not only for partial implementation but also for *unique implementation.*

2 The Model

This study investigates a situation in which a central planner attempts to correctly elicit information from multiple agents. Let $N = \{1, 2, \ldots, n\}$ denote the finite set of all agents, where $n \geq 2$. 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 only 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 [2, 3]:

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

where $t_i \in T_i$ is agent i's type, $\theta_i : \Omega \times T_i \to \{0, 1\}$ represents agent i's honesty, and $\pi_i : \Omega \times T_i \to \Delta(T_{-i})$ denotes agent i's belief about the other agents' types.[4] Each agent i knows her type t_i and the state ω, but does not know the other agents' types t_{-i}. Agent i expects that the other agents' types are distributed according to a probability measure $\pi_i(\omega, t_i) \in \Delta(T_{-i})$. Each agent is either selfish or honest: agent i is selfish (honest) if $\theta_i(\omega, t_i) = 0$ ($\theta_i(\omega, t_i) = 1$, respectively). More details will be subsequently explained.

[3] A number of studies extended the scoring rule to a setting in which a central planner collects information from a group of agents (e.g., [5, 6, 8, 9]). 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).

[4] We denote by $\Delta(Z)$ the space of probability measures on the Borel field of a measurable space Z. We denote $Z \equiv \times_{i \in N} Z_i$, $Z_{-i} \equiv \times_{j \neq i} Z_j$, $z = (z_i)_{i \in N} \in Z$, and $z_{-i} = (z_j)_{j \neq i} \in Z_{-i}$.

The central planner designs a mechanism $G \equiv (M, x)$, where $M = \times_{i \in N} M_i$ denotes a message space, $x = (x_i)_{i \in N}$ 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,

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

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, her payoff is equal to her monetary payoff:

$$U_i(m; \omega, t_i, G) = x_i(m) \qquad \text{if } \theta_i(\omega, t_i) = 0.$$

In contrast, if agent i is honest, she is motivated not only by monetary interest but also by an intrinsic preference for honesty.

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

where $c_i(m, \omega, t_i, G) \in R$ denotes agent i's 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 $\tilde{m}_i \in M_i$,

$$\begin{aligned} &[\theta_i(\omega, t_i) = 1, m_i(\omega) > \tilde{m}_i(\omega), \text{ and } x_i(\tilde{m}_i, m_{-i}) > x_i(m)] \\ &\Rightarrow [c_i(m, \omega, t_i, G) < c_i(\tilde{m}_i, m_{-i}, \omega, t_i, G)] \,. \end{aligned} \tag{1}$$

Condition (1) implies that any honest agent feels more or less guilty about telling lies that generate more a monetary payoff. 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.

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

$$E\left[U_i(s(\omega, t); \omega, t_i, G) \middle| \, \omega, t_i\right] \geq E\left[U_i(m_i, s_{-i}(\omega, t_{-i}); \omega, t_i, G) \middle| \, \omega, t_i\right].$$

3 The Theorem

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

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

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 $(\omega, t_i) \in \Omega \times T_i$,

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

whereas

$$[\theta_i(\omega, t_i) = 1] \Rightarrow \left[s_i(\omega, t_i)(\omega) = 1 \text{ or } s_i(\omega, t_i)(\omega) > E\left[\frac{\sum_{j \neq i} s_j(\omega, t_j)(\omega)}{n-1} \middle| \omega, t_i \right] \right].$$

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

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

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 $E \subset T \equiv \times_{i \in N} T_i$ an event. For convenience, for each event $E \subset T$, we write

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

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

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

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

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

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

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

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 1. *The truthful strategy profile s^* is the unique BNE in the game associated with the quadratic scoring payment rule x^* if and only if*

$$\bigtimes_{i \in N} V_i^{\infty}(E, \omega) = \emptyset \textit{ for all } \omega \in \Omega.$$

The proof is shown in the full version of the paper. From the definition of common knowledge, the necessary and sufficient condition of the theorem clearly implies that $V_i^{\infty}(E^*(\omega), \omega) = \emptyset$ for all $i \in N$ and $\omega \in \Omega$. The theorem states that *all agents, whether selfish or honest*, will 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 Application: 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 [7] 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. [7] show that this study's theorem provides a new and promising direction to solve this problem.

References

1. Abeler, J., Nosenzo, D., Raymond, C.: Preferences for truth-telling. Econometrica **87**(4), 1115–1153 (2019)
2. Bergemann, D., Morris, S.: Robust mechanism design. Econometrica pp. 1771–1813 (2005)
3. Bergemann, D., Morris, S.E.: An Introduction to Robust Mechanism Design. Foundations and Trends in Microeconomics, Now Publishers Inc (2013)
4. Brier, G.W.: Verification of forecasts expressed in terms of probability. Monthly Weather Review **78**(1), 1–3 (1950)
5. Dasgupta, A., Ghosh, A.: Crowdsourced judgement elicitation with endogenous proficiency. In: Proceedings of the 22nd International Conference on World Wide Web. pp. 319–330 (2013)
6. Kong, Y., Schoenebeck, G.: An information theoretic framework for designing information elicitation mechanisms that reward truth-telling. ACM Transactions on Economics and Computation (TEAC) **7**(1), 1–33 (2019)
7. Matsushima, H., Noda, S.: Mechanism design with blockchain enforcement (2020), working Paper
8. Miller, N., Resnick, P., Zeckhauser, R.: Eliciting informative feedback: The peer-prediction method. Management Science **51**(9), 1359–1373 (2005)
9. Prelec, D.: A bayesian truth serum for subjective data. Science **306**(5695), 462–466 (2004)