The revelation principle fails when the format of each agent's strategy is an action

Wu, Haoyang

Wan-Dou-Miao Research Lab

15 November 2020

The revelation principle fails when the format of each agent's strategy is an action

Haoyang Wu *

Wan-Dou-Miao Research Lab, Room 301, Building 3, 718 WuYi Road, 200051, Shanghai, China.

Abstract

In mechanism design theory, a designer would like to implement a social choice function which specifies her favorite outcome for each possible profile of agents' private types. The revelation principle asserts that if a social choice function can be implemented by a mechanism in equilibrium, then there exists a direct mechanism that can truthfully implement it.

This paper aims to propose a failure of the revelation principle. We point out that in any game the format of each agent's strategy is either an informational message or a realistic action. The main result is that: For any given social choice function, if the mechanism which implements it has action-format strategies, then "*honest and obedient*" will not be the equilibrium of the corresponding direct mechanism. Consequently, the revelation principle fails when the format of each agent's strategy is an action.

Key words: Mechanism design; Revelation principle.

1 Introduction

In the framework of mechanism design theory [1–3], there are one designer and some agents labeled as $1, \cdots, I$. [1] Suppose that the designer would like to implement a social choice function which specifies her favorite outcome for each possible profile of agents' types, and each agent's type is modeled as his privacy. In order to implement a social choice function in equilibrium, the

* Corresponding author.
Email address: 18621753457@163.com (Haoyang Wu).

[1] In this paper, the designer is always denoted as "She", and the agent is denoted as "He".

designer constructs a mechanism which specifies each agent's feasible strategy set (*i.e.*, the allowed actions of each agent) and an outcome function (*i.e.*, a rule for how agents' actions get turned into a social choice).

The revelation principle is an important theorem in mechanism design theory. It asserts that if a social choice function can be implemented by a mechanism in equilibrium, then it is truthfully implementable. So far, there have been several criticisms on the revelation principle: Bester and Strausz [4] pointed out that the revelation principle may fail because of imperfect commitment; Epstein and Peters [5] proposed that the revelation principle fails in situations where several mechanism designers compete against each other. Kephart and Conitzer [6] proposed that when reporting truthfully is costless and misreporting is costly, the revelation principle can fail to hold.

Different from these criticisms, this paper aims to propose another failure of the revelation principle. The paper is organized as follows: Section 2 analyses the distinction of two formats of strategy, Section 3 proposes the main result, *i.e.*, the revelation principle fails when each agent's strategy is action-format. Section 4 draws conclusions. Notations about mechanism design theory are given in Appendix, which are cited from MWG's book [1].

2 Two formats of strategy

Note 1: In any game, the format of each agent's strategy is either an informational message or a realistic action. [2]

Example: Practically, only in some restricted cases (such as chess, war simulation game and so on) can each agent's strategy be described as pure information and represented by an informational message (*e.g.*, the strategy in a war simulation game is message-format, since it contains abstract plans of players). On the other hand, in many realistic cases each agent's strategy cannot be described as pure information but must be described as a realistic action (*e.g.*, the strategy in a real war is action-format, since it contains military actions of armies).

Another interesting example is the auction. At first sight each bidder's bid is pure information and looks like a message-format strategy. However, in

[2] Although Note 1 looks simple, it is not trivial. The reason why we emphasize the two formats of strategy is that the revelation principle does not hold for the case of action-format strategies, as will be discussed in Section 3. For simplification, in the following discussions we simply assume that in any game each agent's strategy is of the same format, *i.e.*, we omit the case in which some agents' strategies are message-format and other agents' strategies are action-format.

many practical cases, only the bid information itself is not enough to be a full strategy in the auction. Besides announcing a message-format bid, each bidder must perform a realistic action (*e.g.*, paying money to the auctioneer) in order to really win the auction. Hence, in many practical cases, an auction is indeed a game with action-format strategies.

Next, we will deeply investigate the two formats of strategy respectively.

2.1 Case 1: Mechanism with message-format strategies

Definition 1: A message-format strategy of an agent in a mechanism is a strategy represented by an informational message. The information contained in the message is the strategy itself, which does not need to be carried out realistically. For example, let us consider a chess game, then each player's message-format strategy is his informational plan about how to play chess.

Definition 2: Given a social choice function f, suppose a mechanism $\Gamma = (S_1, \cdots, S_I, g(\cdot))$ implements it in equilibrium with message-format strategies. To clearly describe the case of message-format strategies, we denote each strategy set S_i as M_i, and each agent i's strategy function as $m_i(\cdot) : \Theta_i \to M_i$, where Θ_i is agent i's type set. The outcome function $g(\cdot)$ is denoted as $g_m(\cdot) : M_1 \times \cdots \times M_I \to X$, where the input parameters are message-format strategies and X is the set of outcomes. Hence, the mechanism Γ is denoted as $\Gamma_m = (M_1, \cdots, M_I, g_m(\cdot))$. The game induced by Γ_m is denoted as G_m, which works in a one-stage manner:
Step 1: By using the strategy function $m_i(\cdot)$, each agent i with private type θ_i send message $m_i(\theta_i)$ to the designer. [3]
Step 2: The mechanism Γ_m yields the outcome $g_m(m_1(\theta_1), \cdots, m_I(\theta_I))$.

Definition 3: Suppose the game G_m has an equilibrium, denoted as $m^*(\cdot) = (m_1^*(\cdot), \cdots, m_I^*(\cdot))$. Consider this equilibrium, there is a direct compound mapping from agents' possible types $\hat{\theta} = (\hat{\theta}_1, \cdots, \hat{\theta}_I) \in \Theta$ into the outcome $g_m(m^*(\hat{\theta}))$, which is equal to $f(\hat{\theta})$. Based on the direct compound mapping, we construct a direct mechanism $\bar{\Gamma}_m = (\Theta_1, \cdots, \Theta_I, g_m(m^*(\cdot)))$.

Definition 4: The direct mechanism $\bar{\Gamma}_m$ induces a *one-stage* direct game $\bar{G}_m$ as follows:
Step 1: Each agent i with private type θ_i individually reports a type $\hat{\theta}_i \in \Theta_i$ to the designer. Here, each agent i does not need to be "*honest*", *i.e.*, $\hat{\theta}_i$ can be different from agent i's private type θ_i.
Step 2: By using the equilibrium strategy functions $m^*(\cdot) = (m_1^*(\cdot), \cdots, m_I^*(\cdot))$,

[3] In the following discussions, in order to clearly specify the private type of each agent i, we denote it as θ_i, and any possible type of agent i is denoted as $\hat{\theta}_i \in \Theta_i$.

the direct mechanism $\bar{\Gamma}_m$ calculates $m^*(\hat{\theta}) = (m_1^*(\hat{\theta}_1), \cdots, m_I^*(\hat{\theta}_I))$, and then yields the outcome $g_m(m^*(\hat{\theta}))$.

Note 2: It should be emphasized that the calculated results are pure information, and hence are message-format too. Actually, only when each agent i's strategy $m_i(\cdot)$ is message-format will the calculated results $m^*(\hat{\theta})$ be *legal* message-format parameters of the outcome function $g_m(\cdot)$.

Note 3: By Definition 3, $m^*(\cdot) = (m_1^*(\cdot), \cdots, m_I^*(\cdot))$ is the equilibrium of the game G_m. Then each agent i with private type θ_i finds $m_i^*(\theta_i)$ to be the optimal choice given that all other agents send $m_{-i}^*(\theta_{-i})$. Therefore, in the direct game $\bar{G}_m$, each agent i will find truth-telling $\hat{\theta}_i = \theta_i$ to be the optimal choice given that the others agents tell the truth $\hat{\theta}_{-i} = \theta_{-i}$, and the final outcome is $g_m(m^*(\theta)) = f(\theta)$ for all $\theta \in \Theta$. Thus, for the case of message-format strategies, truth-telling is a Bayesian Nash equilibrium of the direct game $\bar{G}_m$. Consequently, *the revelation principle holds when each agent's strategy is message-format.*

2.2 Case 2: Mechanism with action-format strategies

Definition 5: An action-format strategy of an agent in a mechanism is a strategy represented by a realistic action, which should be performed by himself practically. For example, let us consider a tennis game, then each player's strategy is his realistic action of playing tennis, but not any informational plan of how to play tennis.

Definition 6: Given a social choice function f, suppose a mechanism $\Gamma = (S_1, \cdots, S_I, g(\cdot))$ implements it in equilibrium with action-format strategies. To clearly describe the case of action-format strategies, we denote each strategy set S_i as A_i, and each agent i's strategy function as $a_i(\cdot) : \Theta_i \to A_i$. The outcome function $g(\cdot)$ is denoted as $g_a(\cdot) : A_1 \times \cdots \times A_I \to X$, where the input parameters are action-format strategies. Hence, the mechanism Γ is denoted as $\Gamma_a = (A_1, \cdots, A_I, g_a(\cdot))$. The game induced by Γ_a is denoted as G_a, which works in a one-stage manner:
Step 1: By using action-format strategy functions $(a_1(\cdot), \cdots, a_I(\cdot))$, agents $1, \cdots, I$ with private types $(\theta_1, \cdots, \theta_I)$ perform the action-format strategies $(a_1(\theta_1), \cdots, a_I(\theta_I))$. [4]
Step 2: The mechanism Γ_a yields the outcome $g_a(a_1(\theta_1), \cdots, a_I(\theta_I))$. [5]

[4] For the case of action-format strategies, the designer *observes* the performance of each agent's action. As a comparison, for the case of message-format strategies, the designer *receives* each agent's message.
[5] If in the mechanism Γ_a some agent i only declares a message about how to perform an action but does not realistically perform it, then this declaration is meaningless.

Definition 7: Suppose the game G_a has an equilibrium, denoted as $a^*(\cdot) = (a_1^*(\cdot), \cdots, a_I^*(\cdot))$. Consider this equilibrium, there is a compound mapping from agents' possible types $\hat{\theta} = (\hat{\theta}_1, \cdots, \hat{\theta}_I) \in \Theta$ into the outcome $g_a(a^*(\hat{\theta}))$, which is equal to $f(\hat{\theta})$. Based on the compound mapping, we construct a direct mechanism $\bar{\Gamma}_a = (\Theta_1, \cdots, \Theta_I, \, g_a(a^*(\cdot))$.

Definition 8: According to Myerson [2], the direct mechanism $\bar{\Gamma}_a$ induces a *multistage* direct game $\bar{G}_a$ as follows:
Step 1: Each agent i with private type θ_i individually reports a type $\hat{\theta}_i \in \Theta_i$. Here each agent i does not need to be "*honest*", *i.e.*, $\hat{\theta}_i$ can be different from agent i's true type θ_i.
Step 2: The designer returns a suggestion to each agent i, here the suggestion is just the message-format description of action $a_i^*(\hat{\theta}_i) \in A_i$. In order to specify the suggestion's format more clearly, we denote the suggestion as $a_i^m(\hat{\theta}_i)$;
Step 3: Each agent i individually performs an action $\hat{a}_i \in A_i$. Here each agent i does not need to be "*obedient*", *i.e.*, $\hat{a}_i$ can be different from $a_i^m(\hat{\theta}_i)$.
Step 4: After observing that all actions $\hat{a}_1, \cdots, \hat{a}_I$ have been performed, the direct mechanism $\bar{\Gamma}_a$ yields the outcome $g_a(\hat{a}_1, \cdots, \hat{a}_I)$.

Note 4: Consider Step 1 in Definition 8, each agent reports a type to the designer, either honestly or dishonestly. Note that choosing to be honest or dishonest is each agent's private choice, and cannot be verified by the designer. Hence, no agent is willing to disclose his privacy unless truth-telling is his strictly optimal choice. Put differently, *if the outcome of truth-telling is not superior but only equivalent to the outcome of false-telling, then each agent will certainly prefer false-telling*, because false-telling always protects his privacy.

Note 5: Consider Step 3 in Definition 8, after receiving the designer's suggestion each agent performs an action, either obediently or disobediently. Note that choosing to be obedient or disobedient is each agent's open choice and can be directly observed by the designer. However, in the framework of mechanism design theory, the designer is not a dictator, *i.e.* she can neither control any agent's decision nor punish any agent. Therefore, no agent is willing to be disobedient unless obeying the designer's suggestion is his strictly optimal choice.

3 Main results

Consider the multistage direct game $\bar{G}_a$ induced by the direct mechanism $\bar{\Gamma}_a$ described in Definition 8. According to Myerson [2], the strategy "*honest and obedient*" is the Bayesian Nash equilibrium of the game $\bar{G}_a$: *i.e.*, each agent i not only honestly discloses his private type in Step 1 (*i.e.*, $\hat{\theta}_i = \theta_i$), but also obeys the designer's suggestion in Step 3 (*i.e.*, $\hat{a}_i = a_i^m(\theta_i)$). However, in this section we will point out that Myerson's conclusion will not hold when each agent's strategy is of an action format.

Proposition 1: For a given social choice function $f(\cdot) : \Theta \to X$, suppose there is a mechanism that implements it in Bayesian Nash equilibrium, in which each agent's strategy is of an action format. Then f will not be truthfully implementable, *i.e.*, in the multistage direct game induced by the corresponding direct mechanism, "*honest and obedient*" will no longer be the equilibrium strategy.

Proof: Suppose the mechanism $\Gamma_a = (A_1, \cdots, A_I, g_a(\cdot))$ implements the social choice function $f(\cdot) : \Theta \to X$ in Bayesian Nash equilibrium, in which each agent's strategy is of an action format. By Definition 6 it induces a one-stage game G_a, the equilibrium of which is denoted as $a^*(\cdot) = (a_1^*(\cdot), \cdots, a_I^*(\cdot))$, and $g_a(a^*(\theta)) = f(\theta)$ for all $\theta \in \Theta$. Thus, there is a corresponding direct mechanism $\bar{\Gamma}_a = (\Theta_1, \cdots, \Theta_I,\ g_a(a^*(\cdot)))$ given by Definition 7 and a direct game $\bar{G}_a$ given by Definition 8. In Step 1 of the direct game $\bar{G}_a$, there are different cases for agents $1, \cdots, I$.

Case 1: Each agent is honest
Consider each agent i chooses to be "*honest*" in Step 1 of $\bar{G}_a$ (*i.e.*, $\hat{\theta}_i = \theta_i$), then in Step 2, the suggestion to each agent i will be $a_i^m(\theta_i)$. Since the equilibrium of G_a is $a^*(\cdot) = (a_1^*(\cdot), \cdots, a_I^*(\cdot))$, then the optimal choice of each agent i in Step 3 is to be "*obedient*" (*i.e.*, obeying the suggestion $a_i^m(\theta_i)$ and performing the action $a_i^*(\theta_i)$), given that the others also choose to be "*obedient*". In Step 4, the final outcome will be $g_a(a^*(\theta)) = f(\theta)$ for all $\theta \in \Theta$.

Case 2: One agent is dishonest and the others are honest
Consider there is one agent i that wants to protect his privacy and chooses to be "*dishonest*", and the others still choose to be "*honest*" in Step 1 (*i.e.*, $\hat{\theta}_i \neq \theta_i$, $\hat{\theta}_{-i} = \theta_{-i}$). Then in Step 2, the suggestion to agent i will be $a_i^m(\hat{\theta}_i) \neq a_i^m(\theta_i)$, and the suggestions to the others will still be $a_{-i}^m(\theta_{-i})$.
Since the equilibrium of G_a is $a^*(\cdot) = (a_1^*(\cdot), \cdots, a_I^*(\cdot))$, then the optimal choice of agent i in Step 3 will be "*disobedient*" (*i.e.*, not obeying the suggestion $a_i^m(\hat{\theta}_i)$ but still performing the action $a_i^*(\theta_i)$). The optimal choices of other agents will be "*obedient*" (*i.e.*, performing $a_{-i}^*(\theta_{-i})$).
By Note 5, although in Step 4 the designer can find that agent i is disobedient, she cannot punish him. The direct mechanism $\bar{\Gamma}_a$ will yield the outcome $g_a(a^*(\theta)) = f(\theta)$ for all $\theta \in \Theta$.

It can seen from Case 1 and Case 2 that no matter whether agent i chooses to be "*honest*" or "*dishonest*" in Step 1 of the direct game $\bar{G}_a$, he will always perform the action $a_i^*(\theta_i)$ in Step 3, because everyone knows the equilibrium of G_a is $a^*(\cdot) = (a_1^*(\cdot), \cdots, a_I^*(\cdot))$. Obviously, compared with "*honest*", "*dishonest*" will be strictly beneficial to agent i, since he always protects his private type but obtains the same outcome $g_a(a^*(\theta))$ as yielded in Case 1. Furthermore, Case 2 can be generalized to everyone as follows.

Case 3: Each agent is dishonest
Consider each agent i chooses to be "*dishonest*" in Step 1 (*i.e.* reporting a false type $\hat{\theta}_i \neq \theta_i$), and then chooses to be "*disobedient*" in Step 3 (*i.e.*, not obeying the suggestion $a_i^m(\hat{\theta}_i)$ but performing the action $a_i^*(\theta_i)$). Thus, the final outcome is $g_a(a^*(\theta)) = f(\theta)$ for all $\theta \in \Theta$. As a result, f is *dishonestly* implemented by the direct mechanism $\bar{\Gamma}_a = (\Theta_1, \cdots, \Theta_I, g_a(a^*(\cdot)))$. Note that in Case 3 each agent's private type is protected. Thus, Case 3 is strictly attractive to each agent.

To sum up, when the format of each agent's strategy is an action, "*honest and obedient*" will no longer be the equilibrium strategy of the corresponding direct mechanism $\bar{\Gamma}_a = (\Theta_1, \cdots, \Theta_I, g_a(a^*(\cdot)))$. Hence the revelation principle does not hold when each agent's strategy is action-format. $\square$

4 Conclusions

In this paper, we propose that in any game there are two formats of strategy (*i.e.*, an informational message or a realistic action). In Section 2.1 we point out that the revelation principle holds when each agent's strategy is of message format. However, when each agent's strategy is of an action format, in the multistage direct game induced by the direct mechanism, "*honest and obedient*" will no longer be the Bayesian Nash equilibrium. Therefore, the revelation principle fails when each agent's strategy is of action format.

Appendix: Notations and Definitions

Let us consider a setting with one designer and I agents indexed by $i = 1, \cdots, I$. Each agent i privately observes his *type* θ_i that determines his preference over elements in an outcome set X. The set of possible types for agent i is denoted as Θ_i. The vector of agents' types $\theta = (\theta_1, \cdots, \theta_I)$ is drawn from set $\Theta = \Theta_1 \times \cdots \times \Theta_I$ according to probability density $\phi(\cdot)$, and each agent i's *utility function* over the outcome $x \in X$ given his type θ_i is $u_i(x, \theta_i) \in \mathbb{R}$.

A *mechanism* $\Gamma = (S_1, \cdots, S_I, g(\cdot))$ is a collection of I strategy sets $S_1, \cdots, S_I$ and an outcome function $g : S_1 \times \cdots \times S_I \to X$. The mechanism combined with possible types $(\Theta_1, \cdots, \Theta_I)$, the probability density $\phi(\cdot)$ over the possible realizations of $\theta \in \Theta_1 \times \cdots \times \Theta_I$, and utility functions $(u_1, \cdots, u_I)$ defines a Bayesian game of incomplete information. The strategy function of each agent i in the game induced by Γ is a private function $s_i(\theta_i) : \Theta_i \to S_i$. Each strategy set S_i contains agent i's possible strategies. The outcome function $g(\cdot)$ describes the rule for how agents' strategies get turned into a social choice.

A *social choice function* (SCF) is a function $f : \Theta_1 \times \cdots \times \Theta_I \to X$ that, for each possible profile of the agents' types $\theta_1, \cdots, \theta_I$, assigns a collective choice $f(\theta_1, \cdots, \theta_I) \in X$.

A strategy profile $s^*(\cdot) = (s_1^*(\cdot), \cdots, s_I^*(\cdot))$ is a *Bayesian Nash equilibrium* of mechanism $\Gamma = (S_1, \cdots, S_I, g(\cdot))$ if, for all i and all $\theta_i \in \Theta_i$, $\hat{s}_i \in S_i$, there exists

$$E_{\theta_{-i}}[u_i(g(s_i^*(\theta_i), s_{-i}^*(\theta_{-i})), \theta_i)|\theta_i] \geq E_{\theta_{-i}}[u_i(g(\hat{s}_i, s_{-i}^*(\theta_{-i})), \theta_i)|\theta_i].$$

The mechanism $\Gamma = (S_1, \cdots, S_I, g(\cdot))$ *implements the social choice function $f(\cdot)$ in Bayesian Nash equilibrium* if there is a Bayesian Nash equilibrium of Γ, $s^*(\cdot) = (s_1^*(\cdot), \cdots, s_I^*(\cdot))$, such that $g(s^*(\theta)) = f(\theta)$ for all $\theta \in \Theta$.

A *direct mechanism* is a mechanism $\bar{\Gamma} = (\bar{S}_1, \cdots, \bar{S}_I, \bar{g}(\cdot))$ in which $\bar{S}_i = \Theta_i$ for all i and $\bar{g}(\theta) = f(\theta)$ for all $\theta \in \Theta$. [6] The social choice function $f(\cdot)$ is *truthfully implementable in Bayesian Nash equilibrium* (or *Bayesian incentive compatible*) if $\bar{s}_i^*(\theta_i) = \theta_i$ for all $\theta_i \in \Theta_i$ and $i = 1, \cdots, I$ is a Bayesian Nash equilibrium of the direct mechanism $\bar{\Gamma} = (\bar{S}_1, \cdots, \bar{S}_I, \bar{g}(\cdot))$, in which $\bar{S}_i = \Theta_i$, $\bar{g} = f$. That is, if for all $i = 1, \cdots, I$ and all $\theta_i \in \Theta_i$, $\hat{\theta}_i \in \Theta_i$, there exists

$$E_{\theta_{-i}}[u_i(f(\theta_i, \theta_{-i}), \theta_i)|\theta_i] \geq E_{\theta_{-i}}[u_i(f(\hat{\theta}_i, \theta_{-i}), \theta_i)|\theta_i].$$

Proposition 23.D.1 [1]: (*The Revelation Principle for Bayesian Nash Equilibrium*) Suppose that there exists a mechanism $\Gamma = (S_1, \cdots, S_I, g(\cdot))$ that implements the social choice function $f(\cdot)$ in Bayesian Nash equilibrium. Then $f(\cdot)$ is truthfully implementable in Bayesian Nash equilibrium.

Acknowledgment

This paper has been conceived for more than ten years. In order to memorize this invaluable period, here I would like to cite a famous Chinese verse in Song Dynasty: "*Tao Li Chun Feng Yi Bei Jiu, Jiang Hu Ye Yu Shi Nian Deng*". Thanks a lot to my love: Fang, Yue, Xing, Chen.

[6] The bar symbol is used to distinguish the direct mechanism from the indirect mechanism.

References

[1] A. Mas-Colell, M.D. Whinston and J.R. Green, Microeconomic Theory. Oxford University Press, 1995.

[2] R. Myerson, Optimal coordination mechanisms in generalized principal-agent problems. Journal of Mathematical Economics, vol.10, 67-81, 1982.

[3] Y. Narahari et al, Game Theoretic Problems in Network Economics and Mechanism Design Solutions. Springer, 2009.

[4] H. Bester and R. Strausz, Contracting with imperfect commitment and the revelation principle: The single agent case. Econometrica, Vol.69, No.4, 1077-1098, 2001.

[5] Epstein and Peters, A revelation principle for competing mechanisms. Journal of Economic Theory, 88, 119-160 (1999).

[6] A. Kephart and V. Conitzer, The revelation principle for mechanism design with reporting cost. In Proceedings of the ACM Conference on Electronic Commerce (EC), Maastricht, The Netherlands, 2016.