Mechanism Design with Hidden Information*

Thomas J. Rivera

HEC Paris

April, 2016

1 Introduction

In this paper I address the question of whether the equilibrium induced by a particular mechanism is robust when agents have access to additional information that the designer is unaware of. I consider the context of a general mechanism design problem[1] where agents have private information and take actions that cannot be directly controlled. In such a setting, the mechanism designer chooses a communication equilibrium, of the Bayesian game modeling the agents incentives, in order to maximize some social choice function. Then, the communication equilibrium is implemented, with out loss of generality[2], in the following way: first, players report their private information and then, contingent on this information, the designer draws an action profile from the communication equilibrium distribution and reports to each player their component of the realized action profile (and nothing else). The communication equilibrium is then incentive compatible in the sense that it is optimal for each player to report truthfully and play the action suggested to them.

Next I introduce the possibility that players have access to some extraneous signals that are informative with respect to the true state of the world and ask whether the communication equilibrium described above is robust to this new information. Namely, is the communication equilibrium chosen by the *naive* designer – who believes that the players do not receive any such additional information – robust to the introduction to this information, even when it is very imprecise? What I show is that generically a communication equilibrium is robust to additional information of arbitrarily small precision[3] if and only if whenever a player's incentive constraints are binding, then the mechanism reveals more information to that player about the true state of the world than the information structure does. Namely, if under the communication equilibrium a player of a certain type has a binding constraint, then either (1) the action suggested to them perfectly reveals the true state of the world or (2) the signal they receive from the information structure is a *garbling* of

*Preliminary work. Please do not circulate without the author's consent.

[1] Also referred to as a general principle agent problem a la Myerson (1982)

[2] Due to the revelation principle for games of incomplete information (see Forges (1986) and Myerson(1986).

[3] By precision I mean the maximal distance from the prior to posterior beliefs for any signal of any player.

the information regarding the true state of the world contained in the suggested action.[4] Further, I show that generically a communication equilibrium is robust to any information structure of arbitrarily small precision if and only if the mechanism perfectly reveals the true state of the world to any player of some type whose incentive constraints are binding. I then provide implications of these results with respect to the literature on robustness of mechanisms to higher order beliefs (see *e.g.,* Bergemann and Morris (2005) and Oury and Tercieux (2012)) and propose a new class of mechanisms to help overcome this robustness issue.

2 Illustrative Example

In order to clarify this idea, I will introduce the following (very simple) example. Suppose that there are two agents in a firm who produce a good for the principal. Agent 1 (him) designs the good and has proper incentives to do so (*i.e.,* has a single trivial action), but has private information regarding his ability to design; with equal probability he is either a high productivity designer of type θ_H or a low productivity designer θ_L. Agent 2 (her) has no private information, but the principle cannot perfectly control her effort which I assume takes either a high value (H) or a low value (L). The game modeling the incentives is given as follows:

	L	H
θ_L	$(0,0)$	$(1,1)$

	L	H
θ_H	$(1,2)$	$(2,1)$

In such a setting, both players prefer Agent 2 to take action H when the Agent 1 is the low productivity type; a bad design fails if not enough effort is exerted. But, when Agent 1 is of the high productivity type Agent 2 prefers to exert a low amount of effort while Agent 1 would prefer her to exert a high amount; the gains from exerting the extra effort with a good design disproportionately benefit the designer. Assume that the principle always prefers Agent 2 to exert a high amount of effort.

Now, a communication equilibrium of this game consists of two numbers $p_L \in [0,1]$ and $q_L \in [0,1]$ such that whenever the state is θ_L (θ_H) the principle tells Agent 2 to play her action L (H) with probability p_L (q_L) and her action H with probability $1-p_L$ ($1-q_L$). Given that Agent 1 is of type θ_L and θ_H with equal probability we can see that whether Agent 2 plays L or H she receives an expected payoff of 1. Hence, any communication equilibrium (p_L, q_L) is incentive compatible for Agent 2. Further, Agent 1's constraints to truthfully reveal his type can be written as:

$$1 - p_L \geq 1 - q_L \quad \Rightarrow \quad q_L \geq p_L \qquad \text{[Truthfully reporting } \theta_L\text{]}$$

and

$$q_L + 2 \cdot (1 - q_L) \geq p_L + 2 \cdot (1 - p_L) \quad \Rightarrow \quad p_L \geq q_L \qquad \text{[Truthfully reporting } \theta_H\text{]}$$

[4] What I mean by this is that any information about the true state of the world θ_{-i} contained in the extraneous signal s_i that player i receives can be reproduced from the suggested action a_i that they receive; $\mathbb{P}(\theta_{-i}|a_i, s_i) = \mathbb{P}(\theta_{-i}|a_i)$ for all θ_{-i}.

Therefore, any incentive compatible communication equilibrium has $p_L = q_L$ (Agent 2 plays independently of the state). More importantly, we see that the principle can achieve his most preferred outcome when utilizing the communication equilibrium $(p_L, q_L) = (0, 0)$.

Now, we would like to see if the communication equilibrium $(p_L, q_L) = (0, 0)$ is robust to information regarding the state $\theta \in \{\theta_L, \theta_H\}$. Namely, we could imagine that if Agent 1 and Agent 2 work in close proximity that Agent 2 could receive some imperfect signal regarding whether Agent 1 is of type θ_L or θ_H. We could think of this as Agent 2 frequently monitoring Agent 1 when he is working, glancing at his desk to see how much progress he is making, etc. Given that the principle has no idea how frequently Agent 2 monitors Agent 1 he does not know the underlying information structure from which the additional information Agent 2 receives about Agent 1 is drawn. For example, suppose that when Agent 1 is of type θ the monitoring by Agent 2 produces signal $s \in S = \{s_1, s_2\}$ drawn from the probability distribution $\pi(\cdot|\theta)$. In this context I call $\mathcal{I} = (S, \pi)$ the resulting information structure. For example we could consider $\mathcal{I}_\epsilon$ such that signals are drawn according to the following conditional distribution

	θ_L	θ_H
s_1	$\frac{1}{2} + \epsilon$	$\frac{1}{2} - \epsilon$
s_2	$\frac{1}{2} - \epsilon$	$\frac{1}{2} + \epsilon$

For example, under the information structure $\mathcal{I}_\epsilon$, whenever Agent 1 is of type θ_H, then Agent 2 receives the signal s_2 with probability $\frac{1}{2} + \epsilon$.

Now I will show that for any $\epsilon > 0$ the communication equilibrium $(p_L, q_L) = (0, 0)$ is not robust to the information structure $\mathcal{I}_\epsilon$. Namely, I claim that whenever Agent 2 receives the signal s_2 she has a profitable deviation to play her strategy L. To see why this is the case, we simply note that $\mathbb{P}(\theta = \theta_H | s_2) = \frac{1}{2} + \epsilon$ and therefore, by playing R (*i.e.,* obeying the communication equilibrium) Agent 2 receives an expected payoff of $(\frac{1}{2} - \epsilon) \cdot 1 + (\frac{1}{2} + \epsilon) \cdot 1 = 1$ and by playing L she obtains a payoff of $(\frac{1}{2} + \epsilon) \cdot 2 = 1 + 2 \cdot \epsilon$ and therefore she has a profitable deviation from the communication equilibrium $(p_L, q_L) = (0, 0)$ once she has access to the information structure $\mathcal{I}_\epsilon$ for any $\epsilon > 0$. It is therefore in this sense that I say a particular communication equilibrium is not robust to the information structure $\mathcal{I}$; there exists a profitable deviation given the players updated beliefs about the true state of the world when receiving a particular signal.

References

[1] Bergemann, D. and Morris, S. (2005): "Robust Mechanism Design," *Econometrica*, 73, 1521-1534.

[2] Forges, F. (1986): "An Approach to Communication Equilibria," *Econometrica*, 54 (6), 1375-1385.

[3] Myerson, R. B. (1982): "Optimal Coordination Mechanisms in Generalized principal-Agent Problems," *Journal of Mathematical Economics*, 10 (1), 67-81.

[4] Myerson, R. B. (1986): “Multistage Games with Communication," *Econometrica*, 54, 323-358.

[5] Oury, M. and Tercieux, O. (2012): “Continuous Implementation," *Econometrica*, 80, 1605-1637.