Informed Principal, Moral Hazard, and Limited Liability*

Teddy Mekonnen †

Current Draft: July 8, 2018‡

Abstract

I consider a moral hazard problem with risk neutral parties, limited liability, and an informed principal. The contractible outcome is correlated to both the principal's private information and the agent's hidden action. In contrast to a model without a privately informed principal or without limited liability, I show that the first-best payoff cannot be implemented by any equilibrium mechanism. Furthermore, limited liability precludes the existence of equilibrium refinements such as (Strongly) Neologism proofness.

1 Introduction

In the standard model of moral hazard, informational asymmetries arise because the principal is unable to monitor the agent's effort choice. Yet, under risk neutrality and unlimited liability, these informational asymmetries do not lead to any distortions; the principal can extract the full surplus by "selling the firm" to the agent.

However, the principal may also have private information on how the agent's effort choices affect the distribution of observable outcomes, i.e., the production technology. For example, an employer/principal likely has private information about the difficulty of a task an employee/agent is hired to accomplish. An employer who knows the task is easy may wish to reveal that information to encourage the agent to work, whereas an employer who knows the task is difficult may prefer not to disclose that information. Such opposing signaling incentives

*I am grateful to Eddie Dekel, Asher Wolinsky, and Bruno Strulovici for their invaluable comments.
†California Institute of Technology, SISL. Contact: mekonnen@caltech.edu
‡First draft: October 27, 2017.

could lead to distortions. Nonetheless, Wagner, Mylovanov, and Tröger (2015) show that when the production technology satisfies a full rank condition, there is an equilibrium in which the principal extracts the full surplus.

In this paper, I consider a principal-agent model with risk neutral parties and limited liability. The principal is privately informed about the production technology which satisfies the full rank condition. I show that there exists no equilibrium in which the principal can extract the full surplus and earn her first-best payoff. In contrast, if the principal had no private information or the agent had unlimited liability, there is an equilibrium in which the principal extracts the full surplus.

Since informed principal problems often suffer from equilibrium multiplicity, I also consider the implications of limited liability on two commonly used equilibrium refinements in the informed principal literature: Strongly Neologism proofness (Mylovanov and Tröger, 2012), and Neologism proofness (Farrell, 1993; Grossman and Perry, 1986). I show that a Strongly Neologism proof refinement does not exist in the model, a contrast to the results of Wagner et al. (2015). Furthermore, I show that even the weaker criterion of Neologism proofness has no bite in the model as it does not exist whenever there are multiple equilibria.

The main model is purposefully kept simple. The principal has either a highly productive or a less productive technology, the agent can either work or shirk, and observable outcomes either succeed or fail. The simple $2 \times 2 \times 2$ model highlights the tension that arises between the different types of the principal in an informed-principal moral hazard setting: a less productive principal never wants to reveal her type to the agent as doing so entails providing the agent with high-powered incentive schemes. In contrast, a highly productive principal would benefit from revealing her type to the agent but credibly revealing her private information requires some form of costly signaling. Equilibrium contracts are determined by the preferences of the highly productive type who, depending on the agent's prior belief, chooses to either separate through costly signaling or to pool with the less productive type. The trade-off between separating and pooling leads to a tractable geometric characterization of the entire equilibrium payoff-set which I then use to prove the no-surplus-extraction result.

I also generalize the no-surplus extraction result to a model in which there are T different types of the principal, L different actions for the agent, N different observable outcomes, and an arbitrarily finite level of limited liability.[1] The impossibility result does not hinge on the intractable task of characterizing the entire equilibrium payoff-set in the $T \times L \times N$ model. Instead, under a sorting assumption, I show that if full-surplus extraction were possible, then

[1]Wages must be at least $\underline{w}$ with $\infty < \underline{w} \leq 0$.

there would exist a "fictitious" informed principal game between two types of the principal. The first one has a comparative advantage in providing the work incentives to the agent while the second type has a comparative advantage in providing the truthfulness incentives to the principal. The first type would then be solely responsible for incentivizing the agent to work through bonuses and penalties. However, penalties cannot be too harsh with limited liability. Thus, the first type is too constrained to be able to extract the entire surplus while also providing the agent with enough incentives.

Several papers have noted the existence of specific equilibrium outcomes in which an informed principal fails to extract all the surplus. Karle, Schumacher, and Staat (2016) show that when the principal's type (private information) and the agent's efforts are complements, separating equilibria involve some types of the principal signaling through incentive schemes that are higher-powered than first-best. In environments with unlimited liability and a production technology that violates the full rank condition, Beaudry (1994) shows that the principal may leave rents to the agent in the form of efficiency wages, and Inderst (2001) establishes that the principal's signaling incentives may result in flat or low-powered incentive schemes. These papers however highlight specific forms of signaling distortions that arise when an informed principal offers spot-contracts. In contrast, I take a mechanism design approach in which the principal makes effort recommendations and offers a menu of contingent payments, and I show that all equilibrium outcomes are distorted away from the first-best.

The mechanism design approach in this paper follows Myerson (1983), Maskin and Tirole (1990, 1992), Mylovanov and Tröger (2012, 2014), and Wagner et al. (2015). This paper is also related to a larger literature on moral hazard with an informed principal such as Jost (1996), Chade and Silvers (2002), Mezzetti and Tsoulouhas (2000), Bénabou and Tirole (2003), Lee and Fong (2017), and Kaya (2010), which feature double moral hazard problems, risk averse agents, common agency problems, information acquisition, or dynamic principal-agent relationships.

The rest of the paper is structured as follows: Section 2 describes the $2 \times 2 \times 2$ model of the principal-agent game. Section 3 characterizes the set of equilibrium payoffs and establishes the impossibility of full-surplus extraction. Section 4 establishes the non-existence of Strongly Neologism and Neologism proof refinements. Section 5 extends the no-surplus extraction result to a general $T \times L \times N$ model. Any proofs skipped from the main text are in the Appendix.

2 Model

A risk neutral principal (she) contracts with a risk neutral agent (he) to perform a certain task. The agent can choose either to shirk, $e = 0$, or to work, $e = 1$. The contracting environment is one of hidden action: the principal cannot directly monitor the agent nor can the agent provide hard evidence of his effort choice.

The only publicly observable/contractible primitive of the model is the *S*uccess or *F*ailure of the undertaken task denoted by $x \in X \triangleq \{S, F\}$. The probability the task succeeds depends on the agent's effort $e \in E \triangleq \{0, 1\}$ and is given by $\Pr(x = S|e) = \mu \times e$. The parameter $\mu \in (0, 1)$ captures the level of the agent's productivity or the difficulty of the task. The productivity parameter can be either high, $\mu = \mu_H$, or low, $\mu = \mu_L$, where $\mu_H > \mu_L$. Henceforth, I will refer to $\theta \in \Theta \triangleq \{L, H\}$ as the principal's type and use μ_θ to denote the productivity of the agent when he is matched with a type θ principal.

Given an outcome $x \in X$ and wage $w \in \mathbb{R}_+$, the principal's payoff is $\nu_x - w$ where ν_x is the revenue from outcome x with $\nu_S > \nu_F = 0$.[2] Similarly, given a wage $w \in \mathbb{R}_+$ and an effort choice $e \in E$, the agent's payoff is $w - ce$ where $c > 0$ is the cost of effort. I assume that the agent has limited liability and that both the principal and the agent have a reservation value of zero. Furthermore, I assume it is efficient for the agent to work for all types of the principal and for all types to contract with the agent:

$$\mu_L \nu_S - c \geq 0. \tag{1}$$

2.1 Full Information Game

As a baseline, consider the principal-agent game in which the principal's type is publicly observable. There is a unique equilibrium in which each type θ of the principal extracts the surplus and earns her first-best payoff $v_\theta^{FB} = \mu_\theta \nu_S - c$. The equilibrium can be implemented by each

[2]The assumption $\nu_F = 0$ is made to simplify notation. It is not a normalization as the principal's reservation value is normalized to 0. However, none of the results are affected if we only assume that $\nu_S > \nu_F$ as long as the efficiency assumption (1) is amended to

$$\mu_L(\nu_S - \nu_F) \geq \max\{c, c - \nu_F\}.$$

type θ offering the agent an outcome contingent payment given by

$$w(\theta, x) = \begin{cases} \frac{c}{\mu_\theta} & \text{if} \quad x = S \\ 0 & \text{if} \quad x = F \end{cases}.$$

2.2 Informed Principal Game

For the rest of the paper, I consider a principal who privately observes her type $\theta \in \Theta$. The agent does not receive any exogenous signal about the principal's type. Instead, he holds a commonly known full support prior $p_0 \in \text{int}(\Delta(\Theta))$.[3]

The principal-agent game is split into two stages: the proposal stage and the continuation game. In the proposal stage, the principal first learns her type privately and then proposes a contract $\mathcal{C}$ that is comprised of (i) a finite set of messages the principal can send to the agent, and (ii) payments that are possibly message and outcome dependent. Upon observing the proposal, the agent updates his belief to a posterior $q \in \Delta(\Theta)$. The proposed contract and the agent's posterior together then define a finite perfect-recall extensive-form continuation game $(\mathcal{C}, q)$ as described in Figure 1.[4]

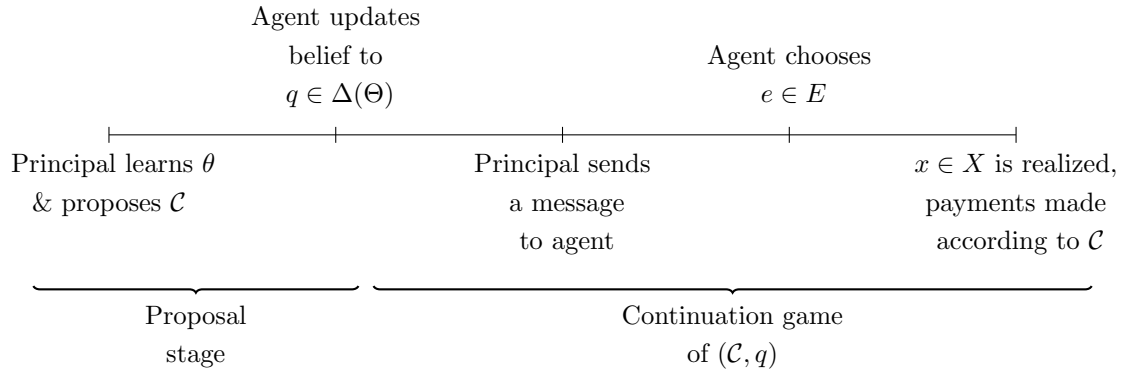


Figure 1: Timing of principal-agent game.

A perfect Bayesian equilibrium (PBE) of the principal-agent game specifies a (possibly random) contract proposal strategy for the principal as well as a posterior belief $q^{\mathcal{C}} \in \Delta(\Theta)$

[3] $\Delta(\Theta)$ represents the space of all probability measures on Θ. For a belief $q \in \Delta(\Theta)$, $\text{supp}(q) \subseteq \Theta$ denotes the support of q. $q \in \text{int}\left(\Delta(\Theta)\right)$ if, and only if, $\text{supp}(q) = \Theta$.

[4] Usually, the agent would also decide to either accept the contract or reject it. However, given limited liability and costless shirking, the agent can guarantee himself at least his reservation value by accepting any contract.

associated with each contract $\mathcal{C}$ such that (i) in any continuation game $(\mathcal{C}, q^{\mathcal{C}})$, the principal's type-dependent messaging strategy and the agent's message-dependent effort strategy constitute a sequential equilibrium, (ii) given a sequential equilibrium outcome in each continuation game, the contract proposal strategy for each type of the principal maximizes her payoff, and (iii) posteriors are derived by Bayes rule whenever possible.

I assume that the players can avail themselves to a public randomization device so that every continuation game has a convex sequential equilibrium payoff set. I further assume that for a given contract $\mathcal{C}$, the correspondence from beliefs q to the sequential equilibrium payoff set of the continuation game $(\mathcal{C}, q)$ is upper-hemicontinuous.[5] Note that contracts are more general than direct revelation mechanisms. While it is without loss of generality to only consider the smaller space of incentive compatible revelation mechanisms on the equilibrium-path, contracts with richer message spaces allow the principal greater flexibility after deviations.

2.3 Feasible Mechanisms

A (direct revelation) mechanism $M \triangleq (r, w)$ is composed of a recommendation policy

$$r : \Theta \to \Delta(E)$$

that maps the principal's report $\hat{\theta} \in \Theta$ to an effort recommendation $\bar{e} \in E$ with probability $r(\bar{e}|\hat{\theta})$, and a compensation policy

$$w : \Theta \times E \times X \to [0, \bar{w}]$$

that maps the report $\hat{\theta} \in \Theta$, the realized recommendation $\bar{e} \in E$, and the observed outcome $x \in X$ to a non-negative wage payment $w(\hat{\theta}, \bar{e}, x)$. The upper bound on wages, $\bar{w}$, is assumed to be large. The restriction is only made to guarantee the existence of an equilibrium but in fact can be dropped. There is no need for random compensation policies as both the principal and agent are risk-neutral.

Suppose the agent obeys the recommendations of a given mechanism M. Then, the principal's payoff from reporting $\hat{\theta}$ when her true type is θ is given by

$$V(\hat{\theta}; \theta, M) = \sum_{\bar{e} \in E} r(\bar{e}|\hat{\theta}) \Big(\mu_\theta \bar{e} \nu_S - \mu_\theta \bar{e} w(\hat{\theta}, \bar{e}, S) - (1 - \mu_\theta \bar{e}) w(\hat{\theta}, \bar{e}, F) \Big).$$

[5] Similar restrictions are employed in Maskin and Tirole (1990, 1992) and Mylovanov and Tröger (2014).

Similarly, suppose the principal reports her type truthfully. Then, the agent's payoff when he holds belief $q \in \Delta(\Theta)$ and employs effort strategy $\xi : E \to E$, a map from recommendation $\bar{e}$ to effort $\xi(\bar{e})$, is given by

$$U(\xi; q, M) = \sum_{\theta \in \Theta} \sum_{\bar{e} \in E} q(\theta) r(\bar{e}|\theta) \Big(\mu_\theta \xi(\bar{e})\, w(\theta, \bar{e}, S) + (1 - \mu_\theta \xi(\bar{e})) w(\theta, \bar{e}, F) - c\, \xi(\bar{e}) \Big).$$

When the principal reports her type truthfully and the agent obeys the effort recommendations, I simplify the notation and write $V(\theta, M)$ and $U(q, M)$. Notice that the agent's belief does not directly affect the principal's payoff. Instead, it affects the agent's incentives to obey the mechanism which in turn affects the principal's payoff.

Definition 1 *Given belief $q \in \Delta(\Theta)$, a mechanism M is q-feasible if it is (i) incentive compatible for the agent to obey recommendations, (ii) incentive compatible for the principal to report her type truthfully, and (iii) individually rational for each type of the principal:*[6]

i. $U(q, M) \geq U(\xi; q, M), \ \ \forall \xi : E \to E.$

ii. $V(\theta, M) \geq V(\hat{\theta}; \theta, M), \ \ \forall \hat{\theta}, \theta \in \Theta.$

iii. $V(\theta, M) \geq 0, \ \ \forall \theta \in \Theta.$

The previous definition of a PBE can now be simplified. First, by the revelation principle, a sequential equilibrium outcome of any continuation game $(\mathcal{C}, q)$ is implementable by some q-feasible mechanism. Second, by the principle of Inscrutability (Myerson, 1983), any on-path outcome of the principal-agent game can be implemented by all types proposing the same direct revelation mechanism.[7] Consequently, the agent cannot infer any new information about the principal's type from the proposed mechanism.

Definition 2 *A mechanism M is a PBE mechanism if (i) M is p_0-feasible, and (ii) for any contract $\tilde{\mathcal{C}}$, there exists a belief $q^{\tilde{\mathcal{C}}}$ and a sequential equilibrium of the continuation game $(\tilde{\mathcal{C}}, q^{\tilde{\mathcal{C}}})$ implemented by a $q^{\tilde{\mathcal{C}}}$-feasible mechanism $\tilde{M}$ such that $V(\theta, M) \geq V(\theta, \tilde{M})$ for all $\theta \in \Theta$.*

[6] Given limited liability, costless shirking, and a zero reservation value, incentive compatibility for the agent implies individual rationality.

[7] The principle of inscrutability does not imply all types will offer the same wages. Instead, the principal (regardless of type) proposes the same "menu" mechanism that specifies possibly different wage schemes for each type report. In the continuation stage of the game, each type then makes a report to select the preferred wage scheme from the menu.

In the next section, I characterize the set of payoffs that are implementable in equilibrium, provide a comparative statics result on the equilibrium payoff set as a function of the agent's prior, and establish the impossibility of implementing the first-best payoffs in any equilibrium.

3 Equilibrium

To make the equilibrium analysis more tractable, I first simplify the space of feasible mechanisms: it is without loss of generality to focus on mechanisms that only recommend work ($\bar{e} = 1$). From the principal's perspective, asking the agent not to work is equivalent to asking the agent to work and giving him the entire surplus.

Lemma 1 *Fix a belief $q \in \Delta(\Theta)$. For any q-feasible mechanism $M \triangleq (r, w)$, there exists another q-feasible mechanism $\tilde{M} \triangleq (\tilde{r}, \tilde{w})$ such that for all $\theta \in \Theta$,*

i. $\tilde{r}(1|\theta) = 1$, *and*

ii. $V(\theta, M) = V(\theta, \tilde{M})$.

Given Lemma 1, I suppress the role of effort recommendations and instead treat mechanisms as a menu of wages, i.e., $M \triangleq \langle w(\theta, x)\rangle_{\theta\in\Theta, x\in X}$. For a given belief $q \in \Delta(\Theta)$, a mechanism $M \triangleq \langle w(\theta, x)\rangle_{\theta\in\Theta, x\in X}$ is q-feasible if it is incentive compatible for the agent to work,

$$\sum_{\Theta} q(\theta)\mu_\theta\big(w(\theta, S) - w(\theta, F)\big) \geq c, \tag{A-IC$_q$}$$

and it is incentive compatible for each type of the principal to report truthfully,

$$\mu_\theta\big(w(\hat{\theta}, S) - w(\theta, S)\big) + (1 - \mu_\theta)\big(w(\hat{\theta}, F) - w(\theta, F)\big) \geq 0 \quad \text{for } \hat{\theta} \neq \theta, \tag{P-IC$_\theta$, $\forall\theta \in \Theta$}$$

and it is individually rational for each type of the principal,

$$\mu_\theta\nu_S - \mu_\theta w(\theta, S) - (1 - \mu_\theta)w(\theta, F) \geq 0. \tag{P-IR$_\theta$, $\forall\theta \in \Theta$}$$

Let $\mathcal{M}(q)$ be the space of q-feasible direct revelation mechanisms that only send work recommendations. Specifically,

$$\mathcal{M}(q) \triangleq \Big\{\langle w(\theta, x)\rangle_{\theta\in\Theta, x\in X} \in [0, \bar{w}]^4 : \langle w(\theta, x)\rangle_{\theta\in\Theta, x\in X} \text{ satisfies A-IC}_q, \text{ P-IC}_\theta, \text{ P-IR}_\theta\ \forall\theta \in \Theta\Big\}.$$

As $\mathcal{M}(q)$ is the intersection of closed half-spaces, it is a convex and compact subspace of $[0,\bar{w}]^4$. Furthermore, $\cap_{q\in\Delta(\Theta)}\mathcal{M}(q)\neq\varnothing$: the mechanism that gives away the firm to the agent with $w(\theta,x)=\nu_x$, $\forall\theta\in\Theta,\forall x\in X$ satisfies all the above constraints for any $q\in\Delta(\Theta)$.

Using Lemma 1, any PBE payoff can be implemented by a mechanism in $\mathcal{M}(p_0)$. However, other mechanisms may lead to the same equilibrium payoff.[8] Hence, it is sometimes more convenient to work with payoffs rather than mechanisms. Let $\mathcal{V}(q)\subseteq\mathbb{R}^2$ be the space of principal-payoff vectors that are implementable by q-feasible mechanisms. Specifically, given a belief $q\in\Delta(\Theta)$, a payoff vector $v\in\mathcal{V}(q)$ if, and only if, there exists a q-feasible mechanism M such that $v=(v_L,v_H)=\Big(V(L,M),V(H,M)\Big)$. As $\mathcal{M}(q)$ is a convex and compact subset of $[0,\bar{w}]^4$ and payoffs are linear and continuous in wages, $\mathcal{V}(q)$ is also convex and compact.

In order to characterize the set of PBE payoff vectors, consider a lower bound on type θ's payoff given by

$$\max_{M}\ V(\theta,M)\quad s.t.\quad M\in\bigcap_{q\in\Delta(\Theta)}\mathcal{M}(q). \tag{2-θ}$$

A mechanism $M\in\bigcap_{q\in\Delta(\Theta)}\mathcal{M}(q)$ is feasible regardless of the agent's belief.[9] Any type of the principal can propose her most preferred mechanism in $\bigcap_{q\in\Delta(\Theta)}\mathcal{M}(q)$ and earn the payoff associated with it regardless of the agent's beliefs. Thus, each type θ's payoff in any equilibrium must be at least as much as the payoff she can earn from the mechanism that solves (2-θ).

The solution to (2-θ) is a mechanism $M^{RSW}\triangleq\langle w(\theta,x)\rangle_{\theta\in\Theta,x\in X}$ with

$w(\theta,x)$	$x=F$	$x=S$
$\theta=L$	0	$\frac{c}{\mu_L}$
$\theta=H$	$\frac{c(\mu_H-\mu_L)}{\mu_H}$	$\frac{c(\mu_H-\mu_L)}{\mu_H}+\frac{c}{\mu_H}$

Table 1: RSW Mechanism.

The low type pays the agent a large bonus of $\frac{c}{\mu_L}$ only if he succeeds at the task, whereas the high type offers the agent an outcome-independent base salary of $\frac{c(\mu_H-\mu_L)}{\mu_H}$ along with a small bonus of $\frac{c}{\mu_H}$ whenever the agent succeeds at the task. Borrowing the terminology of Maskin and Tirole (1992), I refer to the payoff attained from (2-θ) as the *Rothchild-Stiglitz-Wilson* payoff,

[8] For example, even if there may be a unique PBE payoff, it can be implemented by a mechanism $M\in\mathcal{M}(p_0)$ but also by mechanisms with randomized payments. Hence, while the payoff is unique, the mechanism is not.
[9] I reformulate (2-θ) as a linear programming problem in the Appendix.

denoted by v^{RSW} such that

$$v_\theta^{RSW} = \begin{cases} \mu_L \nu_S - c & \text{if} \quad \theta = L \\ \mu_H \nu_S - c - \frac{c(\mu_H - \mu_L)}{\mu_H} & \text{if} \quad \theta = H \end{cases}.$$

Let $\mathcal{V}^*(q) \triangleq \left\{v \in \mathcal{V}(q) : v \geq v^{RSW}\right\}$ denote the set of payoffs that are implementable by q-feasible mechanisms and also dominate the RSW payoff. As mentioned above, any PBE payoff must yield each type θ at least v_θ^{RSW}. Otherwise, type θ can profit by deviating to the RSW mechanism in Table 1 regardless of the agent's off-path belief. Therefore, the equilibrium payoff set is a subset of $\mathcal{V}^*(p_0)$.

The following proposition states that the equilibrium payoff set is in fact $\mathcal{V}^*(p_0)$. Furthermore, the equilibrium set expands as the agent becomes more "optimistic" (as the agent's prior places relatively more mass on the high type).

Proposition 1 *Given a prior $p_0 \in \text{int}(\Delta(\Theta))$, a payoff vector $v \in \mathbb{R}^2$ is implementable by a PBE mechanism if, and only if, $v \in \mathcal{V}^*(p_0)$. Furthermore, given any other belief $p_0' \in \Delta(\Theta)$ with $p_0(H) < p_0'(H)$, $\mathcal{V}^*(p_0) \subseteq \mathcal{V}^*(p_0')$.*

Proof. The proof for the necessary condition in the first statement has already been discussed. The proof for the sufficient condition is a modification of Theorem 1 of Maskin and Tirole (1992) and is provided in the Appendix. Here, I only provide a proof of the second statement.

Any $v \in \mathcal{V}^*(p_0)$ is implementable by some p_0-feasible mechanism M, i.e., $V(\theta, M) = v_\theta$ for all $\theta \in \Theta$. By Lemma 1, we can restrict attention to $M \triangleq \langle w(\theta, x)\rangle_{\theta\in\Theta, x\in X} \in \mathcal{M}(p_0)$. As M is p_0-feasible, the agent is willing to work:

$$\sum_{\theta\in\Theta} p_0(\theta)\mu_\theta \Big(w(\theta, S) - w(\theta, F)\Big) \geq c. \tag{3}$$

Using $v_L \geq v_L^{RSW}$ and Table 1,

$$c \geq \mu_L \Big(w(L, S) - w(L, F)\Big) + \underbrace{w(L, F)}_{\substack{\geq 0 \\ \text{by limited liability}}} \geq \mu_L \Big(w(L, S) - w(L, F)\Big).$$

For (3) to hold, it is then necessary that

$$\mu_H \Big(w(H, S) - w(H, F)\Big) \geq c.$$

Therefore, for any belief $p_0' \in \Delta(\Theta)$ with $p_0'(H) > p_0(H)$, we have

$$\sum_{\theta\in\Theta} p_0'(\theta)\mu_\theta\Big(w(\theta,S)-w(\theta,F)\Big) \geq \sum_{\theta\in\Theta} p_0(\theta)\mu_\theta\Big(w(\theta,S)-w(\theta,F)\Big) \geq c$$

establishing p_0'-feasibility of M.[10] Hence, the payoff vector $\Big(V(L,M),V(H,M)\Big) = v \in \mathcal{V}^*(p_0')$. ■

Proposition 1 implies that a PBE mechanism always exists: the RSW mechanism. However, similar to signaling games, there could be multiple equilibria. In the next proposition, I provide a geometric characterization of the entire equilibrium payoff set.

The following class of payoffs are useful for the characterization: given a belief $q \in \Delta(\Theta)$, let $v^{pool}(q) = \big(v_L^{pool}(q), v_H^{pool}(q)\big) \in \mathcal{V}(q)$ be the payoff vector given by

$$v_\theta^{pool}(q) = \mu_\theta \nu_S - \frac{\mu_\theta c}{\sum_{\theta'\in\Theta} q(\theta')\mu_{\theta'}}.$$

It can be implemented by the pooling mechanism $M^{pool}(q) \in \mathcal{M}(q)$ with wages given by

$w(\theta,x)$	$x=F$	$x=S$
$\theta = L,H$	0	$\frac{c}{\sum_{\theta'\in\Theta} q(\theta')\mu_{\theta'}}$

Table 2: Pooling Mechanism.

It is straightforward to check that $v^{pool}(q)$ is the highest payoff the principal can get when she is forced to offer a type-independent q-feasible mechanism.[11] Furthermore, $v_\theta^{pool}(q)$ is continuous and strictly increasing in the agent's belief, i.e., $v^{pool}(q') > v^{pool}(q)$ if $q'(H) > q(H)$.

Proposition 2 *There exists a cutoff belief $p^* \in \Delta(\Theta)$ such that for any prior $p_0 \in \text{int}(\Delta(\Theta))$, the equilibrium payoff set $\mathcal{V}^*(p_0)$ is given by*

$$\mathcal{V}^*(p_0) = \begin{cases} \{v^{RSW}\} & \text{if} \quad p_0(H) < p^*(H) \\ \text{conv}\Big(\big\{v^{RSW}, v^{pool}(p_0), v^{pool}(p^*)\big\}\Big) & \text{if} \quad p_0(H) \geq p^*(H) \end{cases},$$

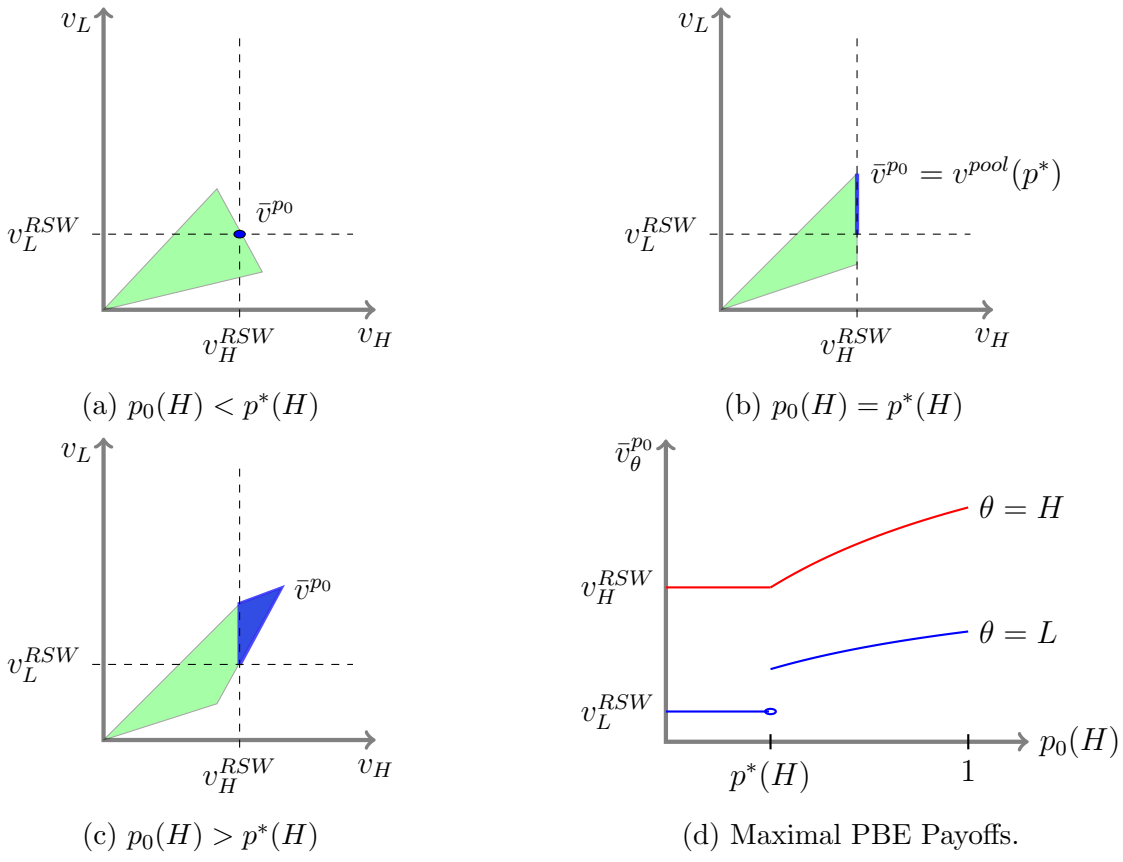
[10] The principal's payoff and her incentives to truthfully report are not directly affected by the agent's beliefs. As long as the agent works, the principal's incentives to truthfully report remain unchanged.

[11] Formally, $v_\theta^{pool} = \max_{M\in\mathcal{M}(q)} \; V(\theta,M) \quad s.t. \; w(L,x) = w(H,x), \forall x \in X$.

where conv$(\cdot)$ *is the convex hull. The belief* p^* *is given by*

$$\big(p^*(L), p^*(H)\big) = \left(\frac{\mu_H}{2\mu_H - \mu_L}, \frac{\mu_H - \mu_L}{2\mu_H - \mu_L}\right).$$

Figure 2a-2c below represent how the equilibrium payoff set $\mathcal{V}^*(p_0)$ changes as a function of the agent's prior. Notice that for $p_0, p_0' \in \text{int}(\Delta(\Theta))$ with $p_0(H) < p_0'(H)$, $\mathcal{V}^*(p_0) \subseteq \mathcal{V}^*(p_0')$.



(a) $p_0(H) < p^*(H)$

(b) $p_0(H) = p^*(H)$

(c) $p_0(H) > p^*(H)$

(d) Maximal PBE Payoffs.

Figure 2: The green area represents $\mathcal{V}(p_0)$, payoffs implementable by p_0-feasible mechanisms. The blue area represents $\mathcal{V}^*(p_0)$, the subset of payoffs implementable by PBE mechanisms.

It is straightforward to see that $v^{RSW} \lneqq v^{pool}(p^*)$.[12] Thus, an immediate consequences of Proposition 2 is that there is a unique PBE payoff vector (namely, v^{RSW}) if, and only if, $p_0(H) < p^*(H)$.

[12]Specifically, $v_L^{RSW} < v_L^{pool}(p^*)$ and $v_H^{RSW} = v_H^{pool}(p^*)$.

Corollary 1 *For any prior $p_0 \in \text{int}(\Delta(\Theta))$, there is a unique Pareto-dominant PBE payoff vector $\bar{v}^{p_0}$, i.e., $\bar{v}^{p_0} \geq v$ for all $v \in \mathcal{V}^*(p_0)$. Furthermore, $\bar{v}^{p_0}$ strictly Pareto dominates all other p_0-feasible payoffs when $p_0(H) > p^*(H)$.*[13]

Proof.

Case 1: $p_0(H) < p^*(H)$.

The result is immediate by setting $\bar{v}^{p_0} = v^{RSW}$.

Case 2: $p_0(H) \geq p^*(H)$.

$v^{pool}(p_0) \geq v^{pool}(p^*) \geq v^{RSW}$. From Proposition 2, $v \in \mathcal{V}^*(p_0) = \text{conv}\Big(\{v^{RSW}, v^{pool}(p_0), v^{pool}(p^*)\} \Big)$. We get the result by setting $\bar{v}^{p_0} = v^{pool}(p_0)$.

Case 3: $p_0(H) > p^*(H)$.

In this case, $v^{pool}(p_0), v^{pool}(p^*)$, and v^{RSW} are not collinear. Thus, $\mathcal{V}^*(p_0)$ has a non-empty interior. Furthermore, we have $v^{pool}(p_0) > v^{pool}(p^*) \geq v^{RSW}$. Thus, $\bar{v}^{p_0} = v^{pool}(p_0) > v$ for all $v \in \mathcal{V}^*(p_0) \backslash \{\bar{v}^{p_0}\}$.

Take any $\tilde{v} \in \mathcal{V}(p_0) \backslash \{\bar{v}^{p_0}\}$. Recall that $\mathcal{V}(p_0)$ is convex. For any $v \in \text{int}(\mathcal{V}^*(p_0)) \subseteq \mathcal{V}(p_0)$, there exists an $\alpha \in (0,1)$ such that $\alpha v + (1-\alpha)\tilde{v} \in \text{int}(\mathcal{V}^*(p_0))$. Since, $\bar{v}^{p_0} > v$, we have that $\alpha \bar{v}^{p_0} + (1-\alpha)\tilde{v} > \alpha v + (1-\alpha)\tilde{v} \geq v^{RSW}$ which implies that $\alpha \bar{v}^{p_0} + (1-\alpha)\tilde{v} \in \mathcal{V}^*(p_0)$. Hence, $\bar{v}^{p_0} > \alpha \bar{v}^{p_0} + (1-\alpha)\tilde{v}$ which implies that $\bar{v}^{p_0} > \tilde{v}$. ■

Henceforth, I refer to $\bar{v}^{p_0}$ as the maximal PBE payoff. In Figure 2a-2c above, the maximal PBE payoff is represented by the north-east vertex of the blue area. Figure 2d depicts how $\bar{v}^{p_0}$ changes as a function of the agent's prior.

To gain some intuition, recall the full information game in Section 2.1. There is a unique equilibrium in which type θ earns her first-best payoff v_θ^{FB} by recommending the agent to work and paying him only a bonus of $\frac{c}{\mu_\theta}$ for successful outcomes. The low type offers the agent a larger bonus than the high type to compensate for her lower productivity.

Now consider the case of incomplete information. The low type prefers pooling with the high type to avoid the large bonus she would otherwise need to offer. In contrast, the high type faces a trade-off between separating and pooling. If the high type separates, she can offer a small bonus but credible separation requires "burning money," e.g., offering a base salary along with a bonus. If the high type pools with the low type, she must offer a bigger bonus as pooling dampens the agent's incentives to work. The more pessimistic the agent, the more his incentives are dampened by pooling and hence, the bigger his bonus needs to be. If the agent

[13] Note that the second statement is about all other payoffs in $\mathcal{V}(p_0)$, not just PBE payoffs in $\mathcal{V}^*(p_0)$.

is too pessimistic, pooling is too costly and the high type prefers the separating mechanism in Table 1. Otherwise, the pooling mechanism in Table 2 is preferable. The cutoff p^* is the belief at which the high type is indifferent between separating and pooling.

Similar to the full information case, with unlimited liability, each type of the informed principal can implement her first-best payoff in equilibrium: Wagner et al. (2015) show that an informed principal can extract the full surplus in equilibrium if the distribution of output satisfies the following full rank condition: there exists a vector $k \in \mathbb{R}^2$ such that

$$\begin{bmatrix} \mu_L e & 1-\mu_L e \\ \mu_H e & 1-\mu_H e \end{bmatrix} \begin{bmatrix} k_1 \\ k_2 \end{bmatrix}$$

is weakly greater than the zero vector with equality for some type $\theta \in \Theta$ when $e = 1$, and strictly less than the zero vector when $e = 0$. The current model satisfies this full rank condition with $k_1 = 1 - \mu_L$ and $k_2 = -\mu_L$. The principal can implement v^{FB} through the mechanism $M^{WMT} = \langle w(\theta, x) \rangle_{\theta \in \Theta, x \in X}$ with[14]

$w(\theta, x)$	$x = F$	$x = S$
$\theta = L$	$c + \frac{c(1-\mu_L)}{p_0(L)\mu_L} - \frac{c}{p_0(L)\mu_L}$	$c + \frac{c(1-\mu_L)}{p_0(L)\mu_L}$
$\theta = H$	c	c

Table 3: Full surplus extraction under unlimited liability.

The high type offers a flat wage while the low type offers an incentive scheme that punishes the agent if he fails at the task. Yet, with limited liability, full surplus extraction is impossible in any equilibrium of the informed principal-game as shown next.

Proposition 3 *For any prior $p_0 \in \text{int}(\Delta(\Theta))$, there exists no equilibrium in which each type of the principal earns her first-best payoff.*

Proof. From Proposition 2, $\mathcal{V}^*(p_0) = \{v^{RSW}\}$ when $p_0(H) < p^*(H)$, and $v^{RSW} \neq v^{FB}$. From Corollary 1, we have $\bar{v}^{p_0} = v^{pool}(p_0) \geq v$ for all $v \in \mathcal{V}^*(p_0)$ when $p_0(H) \geq p^*(H)$. Note that $v_H^{FB} = v_H^{pool}(\delta_H) > v_H^{pool}(p_0)$ where $\delta_\theta \in \Delta(\Theta)$ is the degenerate belief that the principal's type

[14]WMT for Wagner, Mylovanov, and Tröger.

is θ.[15] In either case, $v^{FB} \notin \mathcal{V}^*(p_0)$. ■

Under full information, each type of the principal only needs to provide incentives for the agent to work. In contrast, if the principal has private information, conflicting signaling incentives among the different types ensue; one type of the principal prefers to reveal the private information while another type prefers otherwise. Hence, each type now faces an additional constraint to be truthful.

When the full rank condition holds and there is unlimited liability, the conflicting signaling incentives can be resolved by the different types "trading" the cost of satisfying the agent's work incentive constraint (A-IC) for the cost of satisfying the principal's truthful reporting constraint (P-IC). In particular, the low type of the principal bears the cost of satisfying A-IC (as evidenced in Table 3) because she has a "comparative advantage " in punishing the agent for failure, i.e., $1-\mu_L > 1-\mu_H$. On the other hand, the high type bears the cost of satisfying P-IC as she has no incentives to mimic the low type. However, such a trade involves the low type severely punishing the agent for failed outcomes which is infeasible with limited liability.

4 Refinements

A natural focal point to resolve the multiplicity problem is the unique maximal PBE payoff vector, $\bar{v}^{p_0}$. This section discuss how it relates to two frequently used refinements in the informed principal literature: Strongly Neologism proofness (Mylovanov and Tröger, 2012) and Neologism proofness (Farrell, 1993; Grossman and Perry, 1986).

4.1 Strongly Neologism Proofness

The first refinement I consider is Strongly Neologism proofness (SNP). Given two payoff vectors v and $\tilde{v}$, let $S(v,\tilde{v})$ be the set of types that strictly prefer v to $\tilde{v}$:

$$S(v,\tilde{v}) = \{\theta \in \Theta : v_\theta > \tilde{v}_\theta\}.$$

Definition 3 *A payoff vector $v \in \mathcal{V}(p_0)$ is implementable by a Strongly Neologism proof mechanism if there exists no belief $q \in \Delta(\Theta)$ along with a payoff vector $\tilde{v} \in \mathcal{V}(q)$ satisfying:*

(a) $S(v,\tilde{v}) \cap \text{supp}(q) = \varnothing$, *and*

[15] $\delta_\theta(\hat{\theta}) = 1$ if $\hat{\theta} = \theta$ and 0 otherwise.

(b) $S(\tilde{v}, v) \cap \text{supp}(q) \neq \varnothing$.

In words, a p_0-feasible mechanism M is SNP if there is no deviation mechanism $\tilde{M}$ that is feasible given the agent believes (a) types that strictly prefer M to $\tilde{M}$ do not deviate, and (b) some types that strictly prefer $\tilde{M}$ to M do deviate. The next lemma establishes that if an SNP mechanism exists, then it necessarily implements the maximal PBE payoff vector $\bar{v}^{p_0}$.

Lemma 2 *If a payoff vector $v \in \mathcal{V}(p_0)$ is implementable by a Strongly Neologism proof mechanism, then $v = \bar{v}^{p_0}$.*

Roughly, all types of the principal would (weakly) prefer to swap any PBE mechanism for some mechanism $\bar{M}^{p_0}$ that implements the maximal PBE payoff vector $\bar{v}^{p_0}$. Even if the agent considers $\bar{M}^{p_0}$ an-off path mechanism under some equilibrium, he should recognize that either type of the principal is likely to deviate to it. Therefore, such a deviation should be uninformative of the principal's type.

Proposition 4 *For any prior $p_0 \in \text{int}(\Delta(\Theta))$, there exists no Strongly Neologism Proof mechanism.*

Proof. Fix any prior $p_0 \in \text{int}(\Delta(\Theta))$. Suppose there exists an SNP mechanism. By Lemma 2, the SNP mechanism must implement payoff vector $\bar{v}^{p_0}$.

Take any other belief $q \in \text{int}(\Delta(\Theta))$ such that

$$\max\{p^*(H), p_0(H)\} < q(H) < 1.$$

Such a q exists: full support of the prior implies $p_0(H) < 1$, and $p^*(H) = \frac{\mu_H - \mu_L}{2\mu_H - \mu_L} < 1$. Also,

$$\bar{v}^{p_0} \in \mathcal{V}^*(p_0) \subseteq \mathcal{V}^*(q) \subseteq \mathcal{V}(q)$$

where the first set inclusion follows by applying the comparative statics result in Proposition 1 to $q(H) > p_0(H)$, and the second set inclusion follows by construction. Had the agent's prior been q instead of p_0, the unique maximal q-feasible PBE payoff vector would be $\bar{v}^q$. Furthermore, $\bar{v}^q > \bar{v}^{p_0}$ by Corollary 1 because $q(H) > p^*(H)$.

Therefore, we have (a) $S(\bar{v}^{p_0}, \bar{v}^q) = \varnothing$, and (b) $S(\bar{v}^q, \bar{v}^{p_0}) = \text{supp}(q) = \Theta$. However, this contradicts the supposition that $\bar{v}^{p_0}$ is implementable by an SNP mechanism. ■

Similar to the negative result of Proposition 3, the non-existence result of an SNP mechanism is a direct consequence of limited liability. Under unlimited liability, Wagner et al. (2015) show that the first-best payoff is implementable by an SNP mechanism.

Admittedly, Strongly Neologism proofness is a demanding refinement criterion because it places little structure on the agent's off-path beliefs. For example, consider a deviation from some PBE mechanism implementing payoff $v \in \mathcal{V}^*(p_0)$ to some other off-path contract $\mathcal{C}$. Upon observing the deviation, the agent changes his prior to an off-path belief $q \in \Delta(\Theta)$. The deviation leads to the continuation game $(\mathcal{C}, q)$ which results in the deviation payoff vector $\tilde{v} \in \mathcal{V}(q)$. Suppose the high type strictly prefers the deviation payoff ($\tilde{v}_H > v_H$) while the low type is indifferent ($\tilde{v}_L = v_L$). We would intuitively expect that the high type is more likely to make the deviation. However, upon observing such a deviation, the definition of SNP permits the agent's off-path belief to place almost no mass on the high type, i.e., $q(H) = \epsilon > 0$ with ϵ small, even though such a posterior can only be justified by Bayesian updating if the high type is relatively less likely to deviate.

The next section explores if the additional structure imposed by Neologism proofness fares better as a refinement in the current model. Unfortunately, the answer remains no.

4.2 Neologism Proofness

Similar to SNP, Neologism proofness restricts the agent's off-path belief by placing no mass on the types that strictly lose after a deviation. Additionally, Neologism proofness restricts the agent's off-path belief to place relatively more mass on the types that strictly gain after a deviation.

Definition 4 *A payoff vector $v \in \mathcal{V}(p_0)$ is implementable by a Neologism proof mechanism if there exists no belief $q \in \Delta(\Theta)$ along with a payoff vector $\tilde{v} \in \mathcal{V}(q)$ satisfying:*

(a) $S(v, \tilde{v}) \cap \text{supp}(q) = \varnothing$, *and*

(b) $\frac{q(\theta)}{p_0(\theta)} \geq \frac{q(\theta')}{p_0(\theta')}$ *for all* $\theta \in S(\tilde{v}, v)$ *and all* $\theta' \in \Theta$.

While Neologism proofness is a weaker refinement criteria than SNP, the next proposition shows that it has no bite in the current model– a Neologism proof mechanism does not exist precisely when there are multiple PBE payoff vectors.

Proposition 5 *For any prior $p_0 \in \text{int}(\Delta(\Theta))$ with $p_0(H) \geq p^*(H)$, there exists no Neologism proof mechanism.*

Proof. Suppose a Neologism proof mechanism exists. By the same argument as Lemma 2, it has to implement the maximal PBE payoff vector $\bar{v}^{p_0}$. Fix any prior $p_0 \in \text{int}(\Delta(\Theta))$ with $p_0(H) \geq p^*(H)$. Then, $\bar{v}^{p_0} = v^{pool}(p_0)$ and it can be implemented by the pooling mechanism given in Table 2.

Consider a deviation to a mechanism $\tilde{M} \triangleq \langle \tilde{w}(\theta, x) \rangle_{\theta \in \Theta, x \in X}$ with

$\tilde{w}(\theta, x)$	$x = F$	$x = S$
$\theta = L$	0	$\frac{c}{\sum_{\theta' \in \Theta} p_0(\theta')\mu_{\theta'}}$
$\theta = H$	$\frac{p_0(L)\mu_L}{\sum_{\theta' \in \Theta} p_0(\theta')\mu_{\theta'}} \left(\frac{c(\mu_H - \mu_L)}{\mu_H} \right)$	$\frac{p_0(L)\mu_L}{\sum_{\theta' \in \Theta} p_0(\theta')\mu_{\theta'}} \left(\frac{c(\mu_H - \mu_L)}{\mu_H} \right) + \frac{c}{\mu_H}$

Table 4: Mechanism $\tilde{M}$.

along with a degenerate off-path belief that the principal is the high type, i.e, $q = \delta_H$. We can think of the mechanism $\tilde{M}$ as a random mechanism that implements v^{RSW} with probability $\frac{p_0(L)\mu_L}{\sum_{\theta' \in \Theta} p_0(\theta')\mu_{\theta'}}$ and $v^{pool}(\delta_H)$ with probability $\frac{p_0(H)\mu_H}{\sum_{\theta' \in \Theta} p_0(\theta')\mu_{\theta'}}$, as can be seen in the Figure 3 below.

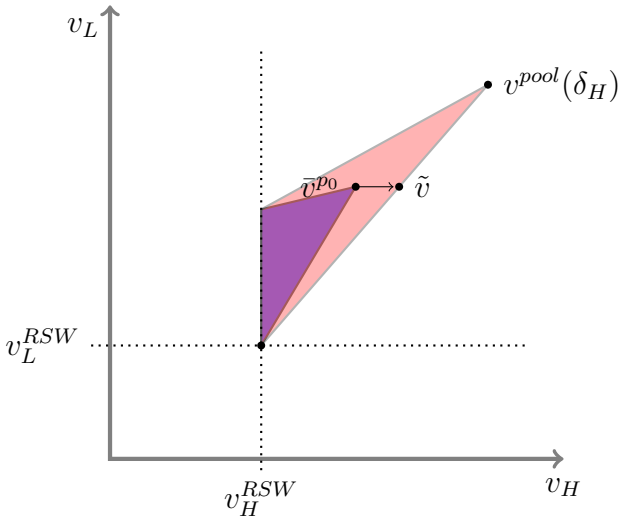


Figure 3: The purple area represents $\mathcal{V}^*(p_0)$ while the pink area represents $\mathcal{V}^*(\delta_H)$. The arrow represents the deviation from $\bar{v}^{p_0}$ to $\tilde{v}$.

The mechanism $\tilde{M}$ is δ_H-feasible: If the principal reports honestly, the agent is willing to obey the recommendation as the payoff difference between working and shirking is

$$\mu_H \Big(\tilde{w}(H, S) - \tilde{w}(H, F) \Big) - c = 0.$$

If the agent is obedient, all types of the principal are willing to report honestly as

$$V(L,\tilde{M}) = \mu_L\nu_S - \frac{c\mu_L}{\sum_{\theta'\in\Theta} p_0(\theta')\mu_{\theta'}} = V(H;L,\tilde{M})$$

and

$$\begin{aligned} V(H,\tilde{M}) =& \mu_H\nu_S - c - \frac{p_0(L)\mu_L}{\sum_{\theta'\in\Theta} p_0(\theta')\mu_{\theta'}}\left(\frac{c(\mu_H-\mu_L)}{\mu_H}\right) \\ >& \mu_H\nu_S - \frac{c\mu_H}{\sum_{\theta'\in\Theta} p_0(\theta')\mu_{\theta'}} \\ =& V(L;H,\tilde{M}). \end{aligned}$$

Hence, the mechanism $\tilde{M}$ implements payoff $\tilde{v} = \Big(V(L,\tilde{M}), V(H,\tilde{M})\Big) \in \mathcal{V}(\delta_H)$.

Furthermore, some simple algebra shows that $\tilde{v}_L = v_L^{pool}(p_0) = \bar{v}_L^{p_0}$ and $\tilde{v}_H > v_H^{pool}(p_0) = \bar{v}_H^{p_0}$. Thus, $S(\bar{v}^{p_0},\tilde{v}) = \varnothing$ and $S(\tilde{v},\bar{v}^{p_0}) = \{H\}$. Since

$$\frac{\delta_H(H)}{p_0(H)} > \frac{\delta_H(L)}{p_0(L)} = 0,$$

we have found a belief δ_H along with a payoff vector $\tilde{v} \in \mathcal{V}(\delta_H)$ that satisfies (a)-(b) of Definition 4. However, this contradicts the supposition that $\bar{v}^{p_0}$ is implementable by a Neologism proof mechanism when $p_0(H) \geq p^*(H)$. ■

5. Generalization

In this section, I consider a T-types$\times L$-actions$\times N$-outcomes model and show a general impossibility result that extends Proposition 3 under a "sorting assumption": for any finite level of limited liability $\underline{w} \in (\infty, 0]$, there exists a set of priors for which full surplus extraction is not possible. Furthermore, this set of priors expands as $\underline{w}$ increases to zero.

Let $\Theta \triangleq \{\theta_1,\theta_2,\ldots,\theta_T\}$ be the different types of the principal, and let $p_0 \in \text{int}(\Delta(\Theta))$ denote the agent's prior. Let $E \triangleq \{e_1,e_2,\ldots,e_L\}$ be the agent's effort space with associated costs $0 = c_1 < c_2 < \ldots < c_L$. Let $\boldsymbol{\nu} = (\nu_n)_{n=1}^N$ be the profit vector for the principal from N outcomes with $\nu_1 < \nu_2 < \ldots < \nu_N$. The outside options for both the agent and the principal are normalized to 0.

Given a principal of type $\theta_t \in \Theta$ and an effort choice $e_l \in E$, let $\boldsymbol{\mu}(\theta_t,e_l) = \big(\mu^n(\theta_t,e_l)\big)_{n=1}^N$ be the probability vector of outcomes (the production function) where $\mu^n(\theta_t,e_l)$ represents the

probability of realizing profit level ν_n. Given a wage vector $\boldsymbol{w} = (w_n)_{n=1}^N$, the principal's payoff is given by $\boldsymbol{\mu}(\theta_t, e_l) \cdot (\boldsymbol{\nu} - \boldsymbol{w})$ while the agent's payoff is given by $\boldsymbol{\mu}(\theta_t, e_l) \cdot \boldsymbol{w} - c_l$. I assume that the principal faces some finite level of limited liability $\underline{w} \in (-\infty, 0]$ so that $w_n \geq \underline{w}$ for $n = 1, 2, \ldots, N$.

Let $S(\theta_t, e_l) = \boldsymbol{\mu}(\theta_t, e_l) \cdot \boldsymbol{\nu} - c_l$ be the surplus generated between a principal of type θ_t and an agent who chooses action e_l. Below, I extend (1) by assuming that there is a single efficient action e_{l^*} (not necessarily the costliest action) for all types of the principal.

A1: Efficiency

There exists some action e_{l^*} with $l^* > 1$ such that for all $\theta_t \in \Theta$,

i. $S_t^* \triangleq S(\theta_t, e_{l^*}) > S(\theta_t, e_l)$ for all $l \neq l^*$, and

ii. $S_t^* \geq 0$.

Consider a (direct revelation) mechanism $M \triangleq \langle \boldsymbol{w}(\theta_t) \rangle_{t=1}^T$ in which each type $\theta_t \in \Theta$ recommends effort e_{l^*} and offers the wage vector $\boldsymbol{w}(\theta_t) \in [\underline{w}, \bar{w}]^N$. If the agent obeys the effort recommendation, the principal's payoff from reporting $\theta_{\hat{t}}$ when her true type is θ_t is given by

$$V(\hat{t}; t, M) = \boldsymbol{\mu}(\theta_t, e_{l^*}) \cdot \big(\boldsymbol{\nu} - \boldsymbol{w}(\theta_{\hat{t}})\big).$$

Similarly, if the principal reports her type truthfully, the agent's payoff when he holds belief $q \in \Delta(\Theta)$ and chooses effort e_l is given by

$$U(l; q, M) = \sum_{t=1}^{T} q(\theta_t) \boldsymbol{\mu}(\theta_t, e_l) \cdot \boldsymbol{w}(\theta_t) - c_l.$$

When the principal reports her type truthfully and the agent obeys the effort recommendation of e_{l^*}, I simplify the notation and write $V(t, M)$ and $U(q, M)$. Given belief $q \in \Delta(\Theta)$, the mechanism M is q-feasible if

$$U(q, M) \geq U(l; q, M), \quad \forall l = 1, 2, \ldots, L, \tag{A-IC$_q^l$}$$

$$U(q, M) \geq 0, \tag{A-IR$_q$}$$

$$V(t, M) \geq V(\hat{t}; t, M), \quad \forall \hat{t}, t = 1, 2, \ldots, T, \text{ and} \tag{P-IC$_t^{\hat{t}}$}$$

$$V(t, M) \geq 0, \quad \forall t = 1, 2, \ldots, T. \tag{P-IR$_t$}$$

The inequalities are the agent's incentive compatibility, agent's individual rationality, the principal's incentive (truthful reporting) compatibility, and the principal's rationality constraints.

In contrast to Section 3, I do not seek to characterize the entire equilibrium payoff set in the $T \times L \times N$ model. Instead, in Lemma 3 below, I provide a necessary and sufficient condition under which first-best surplus extraction is possible by all types of the principal. However, as I show in Proposition 6, the necessary conditions cannot be satisfied for all prior beliefs when a sorting condition holds.

Lemma 3

For a given prior $p_0 \in \text{int}(\Delta(\Theta))$, there exists a first-best equilibrium mechanism M with $V(t, M) = S_t^*$ for each $t = 1, 2, \ldots, T$ if, and only if, there exist vectors $\boldsymbol{k}_t = (k_t^n)_{n=1}^N$ for $t = 1, 2, \ldots, T$ such that

i. $\boldsymbol{\mu}(\theta_t, e_{l^*}) \cdot \boldsymbol{k}_t = 0$ *for each* $t = 1, 2, \ldots, T$,

ii. $\boldsymbol{\mu}(\theta_t, e_{l^*}) \cdot \boldsymbol{k}_{\hat{t}} \geq 0$ *for each* $\hat{t} \neq t$ *and each* $t = 1, 2, \ldots, T$,

iii. $\sum_{t=1}^T p_0(\theta_t) \boldsymbol{\mu}(\theta_t, e_l) \cdot \boldsymbol{k}_t \leq c_l - c_{l^*}$ *for each* $l \neq l^*$, *and*

iv. $k_t^n \geq \underline{w} - c_{l^*}$ *for each* $t = 1, 2, \ldots, T$ *and each* $n = 1, 2, \ldots, N$.

The conditions of Lemma 3 apply for any level of liability including unlimited liability.[16] As a result, these conditions subsume the sufficient full rank conditions in Wagner et al. (2015) for the case of a single efficient action.

In the $2 \times 2 \times 2$ model, (a) working is (weakly) more productive than shirking, regardless of the principal's type, and (b) the high type is (weakly) more productive than the low type, regardless of the agent's action. In the current $T \times L \times N$ model, the production function still lacks structure that would allow us to say much about the equilibrium outcomes. The sorting condition below adds some structure and can be interpreted as establishing a "preferred type" θ_{t^*} and a "preferred action" which coincides with the efficient action e_{l^*}.

A2: Sorting

There exists a type θ_{t^*} such that for some vector $\boldsymbol{y} \in \mathbb{R}^N$,

$$\left[\boldsymbol{\mu}(\theta_{t^*}, e_{l^*}) - \boldsymbol{\mu}(\theta_t, e_{l^*})\right] \cdot \boldsymbol{y} \geq 0, \ \forall t \neq t^* \implies \left[\boldsymbol{\mu}(\theta_{t^*}, e_{l^*}) - \boldsymbol{\mu}(\theta_{t^*}, e_{l'})\right] \cdot \boldsymbol{y} \geq 0 \ \text{ for some } l' < l^*.$$

[16] When $\underline{w} = -\infty$, the last inequality constraints in Lemma 3 are trivially satisfied.

In other words, if it is worthwhile to take the efficient action when working for the preferred type than when working for any other type, then taking the efficient action when working for the preferred type must be at least as good as taking a less costly action.

To gain some intuition, suppose that higher actions and higher types lead to more favorable distributions over outcomes, i.e., $\boldsymbol{\mu}(\theta_t, e_l)$ first-order stochastically dominates $\boldsymbol{\mu}(\theta_{t'}, e_{l'})$ whenever $t > t'$ or $l > l'$. Consider some random outcome that takes values $y_1 \leq y_2 \leq \ldots \leq y_N$. Then, "working" (exerting the efficient level of effort) for the "preferred type" θ_T yields a higher expected outcome than working for any other type, i.e., $\boldsymbol{\mu}(\theta_T, e_{l^*}) \cdot \boldsymbol{y} \geq \boldsymbol{\mu}(\theta_t, e_{l^*}) \cdot \boldsymbol{y}$ for all $t < T$. Furthermore, "working" for the "preferred type" yields a higher expected outcomes than "shirking", i.e., $\boldsymbol{\mu}(\theta_T, e_{l^*}) \cdot \boldsymbol{y} \geq \boldsymbol{\mu}(\theta_T, e_1) \cdot \boldsymbol{y}$. The sorting condition expands the same intuition to the case when $\boldsymbol{y}$ is not an increasing vector.

Example 1: Suppose $N = 2$. The sorting condition is satisfied for $t^* = T$ if $\boldsymbol{\mu}(\theta_t, e_l)$ first-order stochastically dominates $\boldsymbol{\mu}(\theta_{t'}, e_{l'})$ whenever $t > t'$ or $l > l'$. The $2 \times 2 \times 2$ model of Section 3 is a special case of this example.

Example 2: Let $\bar{\boldsymbol{\mu}} = (\bar{\mu}_n)_{n=1}^N$ and $\underline{\boldsymbol{\mu}} = (\underline{\mu}_n)_{n=1}^N$ be two production functions such that $\bar{\boldsymbol{\mu}}$ first-order stochastically dominates $\underline{\boldsymbol{\mu}}$. Let $\lambda : \Theta \times E \to [0, 1]$. Suppose $\boldsymbol{\mu}(\theta_t, e_l) = \lambda(\theta_t, e_l)\bar{\boldsymbol{\mu}} + \big(1 - \lambda(\theta_t, e_l)\big)\underline{\boldsymbol{\mu}}$. The sorting condition is satisfied if $\lambda(\theta_t, e_l)$ increases in t and l. The spanning condition in Grossman and Hart (1983) is a special case of this example in which λ is type-independent.

Proposition 6

Assume (A1) and (A2) hold. For any finite level of limited liability $\underline{w} \in (-\infty, 0]$, there exists a non-empty set of beliefs

$$\mathcal{P}(\underline{w}) \triangleq \left\{ q \in \text{int}(\Delta(\Theta)) : q(\theta_{t^*}) > \frac{c_{l^*-1} - \underline{w}}{c_{l^*} - \underline{w}} \right\}$$

such that for any prior $p_0 \in \mathcal{P}(\underline{w})$, there exists no PBE mechanism M with $V(t, M) = S_t^*$ for each $t = 1, 2, \ldots, T$.

Remark 1: For $\underline{w} < \underline{w}'$, $\mathcal{P}(\underline{w}) \subseteq \mathcal{P}(\underline{w}')$.

Remark 2: For $\underline{w} = c_{l^*-1}$, $\mathcal{P}(\underline{w}) \equiv \text{int}\big(\Delta(\Theta)\big)$. As the model in Section 2 satisfies $\underline{w} = c_1 = 0$, the impossibility result of Proposition 3 is a special case of Proposition 6.

The proof of Proposition 6 first establishes that if first-best surplus extraction is possible, under (A1) and (A2), there necessarily exists a "fictitious" surplus extraction problem with just two types, θ_{t^*} and $\theta_{t'}$ with $t' \neq t^*$. Similar to the $2\times2\times2$ model, types θ_{t^*} and $\theta_{t'}$ "trade" the cost of satisfying the agent's incentive constraint with the cost of satisfying the principal's incentive constraint. Specifically, the preferred type θ_{t^*} has a comparative advantage in satisfying the principal's incentive constraint. Type $\theta_{t'}$ has a comparative advantage in satisfying the agent's incentive compatibility constraint, and thus, bears the entire cost of A-IC$^{l'}_{p_0}$ for some $l' < l^*$.

However, if the agent's prior places a low enough probability on type $\theta_{t'}$, the agent's incentives to choose e_{l^*} over $e_{l'}$ are too weak. Nonetheless, with unlimited liability, the agent's low belief can be off-set by arbitrarily harsh punishments for low profit levels. In contrast, with limited liability, this off-setting is not always possible as the punishments cannot be too harsh. Furthermore, as the lower bound on wages increases, the principal becomes more constrained by how harshly she can punish the agent. Therefore, the set of beliefs for which she cannot extract the first-best-surplus expands.

Proof of Proposition 6. Fix some finite level of limited liability $\underline{w} \in (-\infty, 0]$ and a prior $p_0 \in \text{int}(\Delta(\Theta))$. If there exists a first-best-surplus extracting equilibrium mechanism M, there necessarily exist vectors $\{\boldsymbol{k}_t\}_{t=1}^T$ with $\boldsymbol{k}_t \in \mathbb{R}^N$ that satisfy conditions $i - iv$ of Lemma 3. Using i and ii of Lemma 3, for each $t = 1, 2, \ldots, T$,

$$0 = \boldsymbol{\mu}(\theta_t, e_{l^*}) \cdot (-\boldsymbol{k}_t) \geq \boldsymbol{\mu}(\theta_{\hat{t}}, e_{l^*}) \cdot (-\boldsymbol{k}_t), \; \forall \hat{t} \neq t$$

which, from (A.2), implies that there exists a type t^* such that

$$0 = \boldsymbol{\mu}(\theta_{t^*}, e_{l^*}) \cdot (-\boldsymbol{k}_{t^*}) \geq \boldsymbol{\mu}(\theta_{t^*}, e_{l'}) \cdot (-\boldsymbol{k}_{t^*}) \tag{4}$$

for some $l' < l^*$. Pick some $t' \neq t^*$. The vectors $\boldsymbol{k}_{t^*}$ and $\boldsymbol{k}_{t'}$ satisfy

$$\begin{aligned} c_{l'} - c_{l^*} &\geq \sum_{t=1}^{T} p_0(\theta_t) \boldsymbol{\mu}(\theta_t, e_{l'}) \cdot \boldsymbol{k}_t \\ &\geq \big(1 - p_0(\theta_{t'}) - p_0(\theta_{t^*})\big)(\underline{w} - c_{l^*}) + p_0(\theta_{t'}) \boldsymbol{\mu}(\theta_{t'}, e_{l'}) \cdot \boldsymbol{k}_{t'} + p_0(\theta_{t^*}) \underbrace{\boldsymbol{\mu}(\theta_{t^*}, e_{l'}) \cdot \boldsymbol{k}_{t^*}}_{\geq 0 \text{ by } (4)} \\ &\geq \big(1 - p_0(\theta_{t'}) - p_0(\theta_{t^*})\big)(\underline{w} - c_{l^*}) + p_0(\theta_{t'}) \boldsymbol{\mu}(\theta_{t'}, e_{l'}) \cdot \boldsymbol{k}_{t'} \end{aligned}$$

where the first inequality follows from *iii* of Lemma 3, and the second inequality follows from $k_t^n \geq \underline{w} - c_{l^*}$ by *iv* of Lemma 3. Also from *iv* of Lemma 3, we have $\boldsymbol{\mu}(\theta_{t'}, e_{l'}) \cdot \boldsymbol{k}_{t'} \geq \underline{w} - c_{l^*}$. Therefore,

$$\underline{w} - c_{l^*} \leq \boldsymbol{\mu}(\theta_{t'}, e_{l'}) \cdot \boldsymbol{k}_{t'} \leq \frac{c_{l'} - \underline{w} + \big(p_0(\theta_{t'}) + p_0(\theta_{t^*})\big)(\underline{w} - c_{l^*})}{p_0(\theta_{t'})}$$

which can be rearranged to get $p_0(\theta_{t^*}) \leq \frac{c_{l'} - \underline{w}}{c_{l^*} - \underline{w}} \leq \frac{c_{l^*-1} - \underline{w}}{c_{l^*} - \underline{w}}$.

However, any $p_0 \in \mathcal{P}(\underline{w})$ violates this last inequality, thereby violating the necessary conditions in Lemma 3 for full-surplus extraction. ■

References

Beaudry, P. . Why an informed principal may leave rents to an agent. *International Economic Review*, 35(4):821–832, 1994. ISSN 00206598, 14682354. URL http://www.jstor.org/stable/2526999.

Bénabou, R. and Tirole, J. . Intrinsic and extrinsic motivation. *The Review of Economic Studies*, 70(3):489–520, 2003. ISSN 00346527, 1467937X. URL http://www.jstor.org/stable/3648598.

Chade, H. and Silvers, R. . Informed principal, moral hazard, and the value of a more informative technology. *Economics Letters*, 74(3):291 – 300, 2002. ISSN 0165-1765. URL http://www.sciencedirect.com/science/article/pii/S0165176501005572.

Farrell, J. . Meaning and credibility in cheap-talk games. *Games and Economic Behavior*, 5(4):514 – 531, 1993. ISSN 0899-8256. URL http://www.sciencedirect.com/science/article/pii/S0899825683710298.

Grossman, S. J. and Hart, O. D. . An analysis of the principal-agent problem. *Econometrica: Journal of the Econometric Society*, pages 7–45, 1983.

Grossman, S. J. and Perry, M. . Perfect sequential equilibrium. *Journal of Economic Theory*, 39(1):97 – 119, 1986. ISSN 0022-0531. URL http://www.sciencedirect.com/science/article/pii/0022053186900220.

Inderst, R. . Incentive schemes as a signaling device. *Journal of Economic Behavior & Organization*, 44(4):455 – 465, 2001. ISSN 0167-2681. URL http://www.sciencedirect.com/science/article/pii/S0167268100001426.

Jost, P.-J. . On the role of commitment in a principal–agent relationship with an informed principal. *Journal of Economic Theory*, 68(2):510 – 530, 1996. ISSN 0022-0531. URL http://www.sciencedirect.com/science/article/pii/S0022053196900289.

Karle, H. , Schumacher, H. , and Staat, C. . Signaling quality with increased incentives. *European Economic Review*, 85:8 – 21, 2016. ISSN 0014-2921. URL http://www.sciencedirect.com/science/article/pii/S0014292116300113.

Kaya, A. . When does it pay to get informed? *International Economic Review*, 51(2):533–551, 2010. ISSN 1468-2354. doi: 10.1111/j.1468-2354.2010.00592.x.

Lee, F. X. and Fong, Y.-f. . Signaling by an informed service provider. *Journal of Economics & Management Strategy*, 2017. ISSN 1530-9134. URL http://dx.doi.org/10.1111/jems.12208.

Maskin, E. and Tirole, J. . The principal-agent relationship with an informed principal: The case of private values. *Econometrica*, 58(2):379–409, 1990. ISSN 00129682, 14680262. URL http://www.jstor.org/stable/2938208.

Maskin, E. and Tirole, J. . The principal-agent relationship with an informed principal, ii: Common values. *Econometrica*, 60(1):1–42, 1992. ISSN 00129682, 14680262. URL http://www.jstor.org/stable/2951674.

Mezzetti, C. and Tsoulouhas, T. . Gathering information before signing a contract with a privately informed principal. *International Journal of Industrial Organization*, 18(4):667 – 689, 2000. ISSN 0167-7187. URL http://www.sciencedirect.com/science/article/pii/S016771879800040X.

Myerson, R. B. . Mechanism design by an informed principal. *Econometrica*, 51(6):1767–1797, 1983. ISSN 00129682, 14680262. URL http://www.jstor.org/stable/1912116.

Mylovanov, T. and Tröger, T. . Informed-principal problems in environments with generalized private values. *Theoretical Economics*, 7(3):465–488, 2012. ISSN 1555-7561. URL http://dx.doi.org/10.3982/TE787.

Mylovanov, T. and Tröger, T. . Mechanism design by an informed principal: Private values with transferable utility. *The Review of Economic Studies*, 81(4):1668, 2014. URL +http://dx.doi.org/10.1093/restud/rdu019.

Wagner, C. , Mylovanov, T. , and Tröger, T. . Informed-principal problem with moral hazard, risk neutrality, and no limited liability. *Journal of Economic Theory*, 159:280–289, 2015. URL http://www.sciencedirect.com/science/article/pii/S0022053115000836.

5 Appendix

Proof of Lemma 1. Take any q-feasible mechanism $M \triangleq (r, w)$ and let $v = \Big(V(L, M), V(H, M)\Big)$. Construct a new mechanism $\tilde{M} \triangleq (\tilde{r}, \tilde{w})$ such that for all $\theta \in \Theta$ and $x \in X$,

(i) $\tilde{r}(1|\theta) = 1$,

(ii) $\tilde{w}(\theta, 0, x) = 0$, and

(iii) $\tilde{w}(\theta, 1, x) = r(1|\theta)w(\theta, 1, x) + r(0|\theta)\Big(w(\theta, 0, F) + \nu_x\Big)$.

If the principal reports her type honestly, the agent is willing to follow the work recommendation as his payoff difference from working versus shirking is

$$\sum_{\theta\in\Theta} q(\theta)\left\{\mu_\theta\Big(\tilde{w}(\theta, 1, S) - \tilde{w}(\theta, 1, F)\Big) - c\right\}$$

$$= \underbrace{\sum_{\theta\in\Theta} q(\theta)r(1|\theta)\left\{\mu_\theta\left(w(\theta, 1, S) - w(\theta, 1, F)\right) - c\right\}}_{\substack{\geq 0 \\ \text{by } q\text{-feasibility of } M}} + \sum_{\theta\in\Theta} q(\theta)r(0|\theta)\underbrace{\left\{\mu_\theta\nu_S - c\right\}}_{\substack{\geq 0 \\ \text{by (1)}}} \geq 0.$$

If the agent is obedient, then for all $\theta, \hat{\theta} \in \Theta$

$$\begin{aligned}
V(\hat{\theta}; \theta, \tilde{M}) =& \mu_\theta\nu_S - \mu_\theta\tilde{w}(\hat{\theta}, 1, S) - (1 - \mu_\theta)\tilde{w}(\hat{\theta}, 1, F) \\
=& \mu_\theta\nu_S - \mu_\theta\left\{r(1|\hat{\theta})w(\hat{\theta}, 1, S) + r(0|\hat{\theta})\big[w(\hat{\theta}, 0, F) + \nu_S\big]\right\} \\
& - (1 - \mu_\theta)\left\{r(1|\hat{\theta})w(\hat{\theta}, 1, F) + r(0|\hat{\theta})w(\hat{\theta}, 0, F)\right\} \\
=& \sum_{\bar{e}\in E} r(\bar{e}|\hat{\theta})\left\{\mu_\theta\bar{e}\nu_S - \mu_\theta\bar{e}w(\hat{\theta}, \bar{e}, S) - (1 - \mu_\theta\bar{e})w(\hat{\theta}, \bar{e}, F)\right\} \\
=& V(\hat{\theta}; \theta, M)
\end{aligned}$$

and truthful reporting remains incentive compatible for all types of the principal under $\tilde{M}$. Thus, $\tilde{M}$ satisfies A-IC$_q$ and P-IC$_\theta$ for all $\theta \in \Theta$ and implements the payoff $v \geq 0$, i.e., $\tilde{M}$ is q-feasible. ■

Linear program for RSW mechanism

The problem in (2-θ) can be reformulated as minimizing average wage payments over deterministic wage schemes $W \triangleq \langle W_\theta^x \rangle_{\theta\in\Theta, x\in X} \in [0,\bar{w}]^4$ with $W_\theta^x = w(\theta, x)$. In other words, (2-$\theta$) is equivalent to

$$\min_{W\in[0,\bar{w}]^4} \ \mu_\theta W_\theta^S + (1-\mu_\theta) W_\theta^F \tag{2'-θ}$$

$$s.t. \ \sum_{\theta'\in\Theta} q(\theta')\mu_{\theta'}\Big(W_{\theta'}^S - W_{\theta'}^F\Big) \geq c, \tag{A-IC$_q$, $\forall q \in \Delta(\Theta)$}$$

$$\mu_{\theta'}\Big(W_{\theta''}^S - W_{\theta'}^S\Big) + (1-\mu_{\theta'})\Big(W_{\theta''}^F - W_{\theta'}^F\Big) \geq 0, \ \forall \theta'' \in \Theta \tag{P-IC$_{\theta'}$, $\forall \theta' \in \Theta$}$$

$$\mu_{\theta'}\nu_S - \mu_{\theta'} W_{\theta'}^S - (1-\mu_{\theta'}) W_{\theta'}^F \geq 0. \tag{P-IR$_{\theta'}$, $\forall \theta' \in \Theta$}$$

I relax the above linear program by dropping the P-IR$_\theta$ constraints. Furthermore, notice that if some given wage scheme $W \in [0,\bar{w}]^4$ satisfies both A-IC$_{\delta_H}$ and A-IC$_{\delta_L}$, it also satisfies A-IC$_q$ for all $q \in \Delta(\Theta)$. Let v_θ^{RSW} be the payoff attained from minimizing type θ's expected wage payments in the following relaxed linear program with four constraints:

$$\min_{W\in[0,\bar{w}]^4} \ \mu_\theta W_\theta^S + (1-\mu_\theta) W_\theta^F \tag{2''-θ}$$

$$s.t. \ \mu_{\theta'}\Big(W_{\theta'}^S - W_{\theta'}^F\Big) \geq c, \tag{A-IC$_{\delta_{\theta'}}$, $\forall \theta' \in \Theta$}$$

$$\mu_{\theta'}\Big(W_{\theta''}^S - W_{\theta'}^S\Big) + (1-\mu_{\theta'})\Big(W_{\theta''}^F - W_{\theta'}^F\Big) \geq 0, \ \forall \theta'' \in \Theta. \tag{P-IC$_{\theta'}$, $\forall \theta' \in \Theta$}$$

Notice that the constraint set for program (2''-θ) is independent of the principal's type θ and the agent's belief q. The only difference between (2''-H) and (2''-L) is the objective function we want to minimize. The wage scheme in Table 1 solves both programs, and satisfies all the constraints of (2'-θ).

Proof of Proposition 1. The proof for sufficiency closely follows Theorem 1 of Maskin and Tirole (1992) and is a consequence of the next two lemmas.

Lemma 4 *There exists a belief $q^* \in \text{int}(\Delta(\Theta))$ such that v^{RSW} is q^*-undominated, i.e., there exists no payoff $v \in \mathcal{V}(q^*)$ such that $v_\theta \geq v_\theta^{RSW}$ for all $\theta \in \Theta$ and strictly for at least one type.*

Proof. For some weight $\omega \in (0,1)$, consider the program

$$\min_{W \in [0,\bar{w}]^4} \omega\Big(\mu_L W_L^S + (1-\mu_L)W_L^F\Big) + (1-\omega)\Big(\mu_H W_H^S + (1-\mu_H)W_H^F\Big) \tag{5}$$

$$s.t.\ \ \mu_\theta\Big(W_\theta^S - W_\theta^F\Big) \geq c, \qquad (\text{A-IC}_{\delta_\theta},\ \forall \theta \in \Theta)$$

$$\mu_\theta\Big(W_{\theta'}^S - W_\theta^S\Big) + (1-\mu_\theta)\Big(W_{\theta'}^F - W_\theta^F\Big) \geq 0,\ \forall \theta' \in \Theta. \qquad (\text{P-IC}_\theta,\ \forall \theta \in \Theta)$$

The constraint set of (5) is equivalent to that of (2″-θ). Furthermore, (2″-H) and (2″-L) correspond to the cases $\omega = 0$ and $\omega = 1$ respectively. As the RSW wage scheme from Table 1, denoted by W^{RSW}, solves both (2″-H) and (2″-L), it also solves (5) for any $\omega \in (0,1)$. Additionally, W^{RSW} along with the associated Lagrange multipliers is a strictly complementary solution to the primal-dual problems of (5).

At the wages prescribed by W^{RSW}, both A-IC$_{\delta_H}$ and A-IC$_{\delta_L}$ bind. Let φ_θ^ω be the strictly positive multiplier associated with A-IC$_{\delta_\theta}$ for a given $\omega \in (0,1)$. Construct a full support belief $q^\omega \in \Delta(\Theta)$ such that

$$q^\omega(\theta) = \frac{\varphi_\theta^\omega}{\sum_{\theta' \in \Theta} \varphi_{\theta'}^\omega}.$$

Fix the weight at some $\omega^* \in (0,1)$, and let $q^* \equiv q^{\omega^*}$ and $\varphi_\theta^* \equiv \varphi_\theta^{\omega^*}$. Now consider the program

$$\min_{W \in [0,\bar{w}]^4} \omega^*\Big(\mu_L W_L^S + (1-\mu_L)W_L^F\Big) + (1-\omega^*)\Big(\mu_H W_H^S + (1-\mu_H)W_H^F\Big) \tag{5′}$$

$$s.t.\ \ \sum_{\theta \in \Theta} q^*(\theta)\mu_\theta\Big(W_\theta^S - W_\theta^F\Big) \geq c, \qquad (\text{A-IC}_{q^*})$$

$$\mu_\theta\Big(W_{\theta'}^S - W_\theta^S\Big) + (1-\mu_\theta)\Big(W_{\theta'}^F - W_\theta^F\Big) \geq 0,\ \forall \theta' \in \Theta. \qquad (\text{P-IC}_\theta,\ \forall \theta \in \Theta)$$

A solution to (5′) is a point on the Pareto-frontier of q^*-feasible mechanisms, i.e., a mechanism implementing a q^*-undominated payoff vector. We can see that the Lagrangian of (5) and (5′) coincide at W^{RSW} by setting the multiplier associated with (A-IC$_{q^*}$) to $\sum_{\theta' \in \Theta} \varphi_{\theta'}^*$. Thus, W^{RSW} is also a solution to program (5′) and v^{RSW} is q^*-undominated. ■

Let $\mathcal{V} \triangleq \text{cl}\Big(\text{conv}\Big(\bigcup_{q \in \Delta(\Theta)} \mathcal{V}(q)\Big)\Big)$ be the convex closure of all feasible payoffs. For any payoff

vector $v \in \mathcal{V}$, there is a (random) direct revelation mechanism that can implement v.[17]
Fix a contract $\mathcal{C}$. Let $\Gamma(\mathcal{C}, \cdot) : \Delta(\Theta) \to \mathcal{V}$ be the payoff correspondence so that $\Gamma(\mathcal{C}, q)$ is the set of principal-payoff vectors sustained by a sequential equilibrium of the continuation game $(\mathcal{C}, q)$. As already mentioned in the text, I assume that $\Gamma(\mathcal{C}, q)$ is non-empty, convex, and upper-hemicontinuous. Furthermore, $\Gamma(\mathcal{C}, q) \subseteq \mathcal{V}(q)$ by the revelation principle.
Fix an $\epsilon \in (0, 1)$ and let

$$\mathcal{A}^{\epsilon}_{\theta}(v) = \arg\max_{a \in [\epsilon, 1]} \; a v_{\theta} + (1 - a) v^{RSW}_{\theta},$$

and let $\mathcal{A}^{\epsilon}(v) \triangleq \mathcal{A}^{\epsilon}_{L}(v) \times \mathcal{A}^{\epsilon}_{H}(v)$. A vector $(\alpha_L, \alpha_H) = \alpha \in \mathcal{A}^{\epsilon}(v)$ gives the probabilities with which each type of the principal chooses payoff vector v over v^{RSW} with the constraint that each type must choose v at least with probability $\epsilon > 0$.

Let $\mathcal{Q}^{\epsilon} : [\epsilon, 1]^2 \to \Delta(\Theta)$ be a mapping from a given choice probability vector $\alpha \in [\epsilon, 1]^2$ and the belief q^* from Lemma 4 to a Bayes-updated posterior belief $\mathcal{Q}^{\epsilon}(\alpha) \in \text{int}(\Delta(\Theta))$ with

$$\mathcal{Q}^{\epsilon}(\theta; \alpha) = \frac{q^*(\theta)\alpha_{\theta}}{\sum_{\theta' \in \Theta} q^*(\theta')\alpha_{\theta'}}.$$

Define the correspondence $T^{\epsilon}_{\mathcal{C}}$ that maps $\mathcal{V} \times [\epsilon, 1]^2 \times \Delta(\Theta)$ to itself:

$$T^{\epsilon}_{\mathcal{C}}(v, \alpha, q) = \Gamma(\mathcal{C}, q) \times \mathcal{A}^{\epsilon}(v) \times \mathcal{Q}^{\epsilon}(\alpha).$$

The correspondence $T^{\epsilon}_{\mathcal{C}}$ is upper-hemicontinuous, convex-valued, and closed. Therefore, it has a fixed point, $(v^{\epsilon}, \alpha^{\epsilon}, q^{\epsilon})$.
For intuition, consider the following iterative process for a given contract $\mathcal{C}$: Pick an arbitrary belief $q^1 \in \Delta(\Theta)$ which defines the continuation game $(\mathcal{C}, q^1)$. Pick a sequential equilibrium payoff vector $v^1 \in \Gamma(\mathcal{C}, q^1)$. Pick a choice probability $\alpha^1 \in \mathcal{A}^{\epsilon}(v^1)$ which describes the principal's optimal behavior when choosing between v^1 and v^{RSW}. Based on this behavior, the agent updates his belief from q^* to $q^2 = \mathcal{Q}^{\epsilon}(\alpha^1) \in \Delta(\Theta)$. We now have a different continuation game $(\mathcal{C}, q^2)$. Pick a new sequential equilibrium payoff vector $v^2 \in \Gamma(\mathcal{C}, q^2)$ and a new probability $\alpha^2 \in \mathcal{A}^{\epsilon}(v^2)$ of choosing v^2 over v^{RSW}. Based on this behavior, the agent updates his belief from q^* to $q^3 = \mathcal{Q}^{\epsilon}(\alpha^2) \in \Delta(\Theta)$. And so on.
The fixed point $(v^{\epsilon}, \alpha^{\epsilon}, q^{\epsilon})$ is interpreted as follows: each type θ deviates from the RSW mechanism to the contract $\mathcal{C}$ with probability $\alpha^{\epsilon}_{\theta} \geq \epsilon > 0$. Based on this behavior, the agent updates his belief from q^* to q^{ϵ}. This results in the continuation game $(\mathcal{C}, q^{\epsilon})$ and the subsequent se-

[17] If the mechanism is random, we can use the public randomization device to coordinate beliefs.

quential equilibrium outcome of the continuation game $v^\epsilon \in \Gamma(\mathcal{C}, q^\epsilon)$. Based on this outcome, type θ's constrained choice of deviating from the RSW mechanism to $\mathcal{C}$ is optimal.

Lemma 5 *For any contract $\mathcal{C}$, there exists a belief $q \in \Delta(\Theta)$ and a payoff vector $v \in \Gamma(\mathcal{C}, q)$ such that $v^{RSW} \geq v$.*

Proof. Suppose not! Then, there exists a contract $\mathcal{C}$ such that for all beliefs $q \in \Delta(\Theta)$ and all payoffs $v \in \Gamma(\mathcal{C}, q)$, some type of the principal strictly prefers v to v^{RSW}: $S(v, v^{RSW}) \neq \varnothing$.[18] Construct a fixed point $(v^\epsilon, \alpha^\epsilon, q^\epsilon)$ for $\epsilon > 0$ as described above and let

$$(v^0, \alpha^0, q^0) = \lim_{\epsilon \to 0} (v^\epsilon, \alpha^\epsilon, q^\epsilon).$$

The limit is well defined: $\forall q \in \Delta(\Theta)$ and $\forall v \in \Gamma(\mathcal{C}, q)$, $S(v, v^{RSW}) \neq \varnothing$ implies

$$0 \notin \lim_{\epsilon \to 0} \bigcup_{q \in \Delta(\Theta)} \bigcup_{v \in \Gamma(\mathcal{C}, q)} \mathcal{A}^\epsilon(v).$$

Thus, q^0 is a well-defined probability distribution over Θ . As $v^0 \in \Gamma(\mathcal{C}, q^0) \subseteq \mathcal{V}(q^0)$, it is implementable by a q^0-feasible mechanism that always recommends work. Let $W^0 = \langle W_\theta^{0,x} \rangle_{\theta \in \Theta, x \in X}$ be the wage scheme associated with such a q^0-feasible mechanism.
Construct a new wage scheme $\tilde{W} = \langle \tilde{W}_\theta^x \rangle_{\theta \in \Theta, x \in X}$ with $\tilde{W}_\theta^x = \alpha_\theta^0 W_\theta^{0,x} + (1 - \alpha_\theta^0) W_\theta^{RSW,x}$. The new wage scheme is q^*-feasible. The agent is willing to work as A-IC$_{q^*}$ is satisfied:

$$\sum_{\theta \in \Theta} q^*(\theta) \mu_\theta \left(\tilde{W}_\theta^S - \tilde{W}_\theta^F \right)$$

$$= \left(\sum_{\theta' \in \Theta} q^*(\theta') \alpha_{\theta'}^0 \right) \underbrace{\sum_{\theta \in \Theta} \underbrace{\frac{q^*(\theta) \alpha_\theta^0}{\sum_{\theta' \in \Theta} q^*(\theta') \alpha_{\theta'}^0}}_{= q^0(\theta)} \mu_\theta \left(W_\theta^{0,S} - W_\theta^{0,F} \right)}_{\substack{\geq c \\ \text{by } q^0\text{-feasibility of } W^0}} +$$

$$\left(\sum_{\theta' \in \Theta} q^*(\theta') (1 - \alpha_{\theta'}^0) \right) \underbrace{\sum_{\theta \in \Theta} \frac{q^*(\theta)(1 - \alpha_\theta^0)}{\sum_{\theta' \in \Theta} q^*(\theta')(1 - \alpha_{\theta'}^0)} \mu_\theta \left(W_\theta^{RSW,S} - W_\theta^{RSW,F} \right)}_{\substack{\geq c \\ \text{by } q\text{-feasibility of } W^{RSW} \ \forall q \in \Delta(\Theta)}} \geq c.$$

[18] $S(\tilde{v}, v) = \{\theta \in \Theta : \tilde{v}_\theta > v_\theta\}$ as defined in Section 4.1.

The principal is willing to report truthfully as P-IC$_\theta$ is satisfied for all $\theta \in \Theta$:

$$\begin{aligned}
&\mu_\theta \left(\tilde{W}^S_{\theta'} - \tilde{W}^S_\theta\right) + (1-\mu_\theta)\left(\tilde{W}^F_{\theta'} - \tilde{W}^F_\theta\right) \\
&= \alpha^0_{\theta'} \underbrace{\left[\mu_\theta W^{0,S}_{\theta'} + (1-\mu_\theta) W^{0,F}_{\theta'}\right]}_{\geq \mu_\theta W^{0,S}_\theta + (1-\mu_\theta) W^{0,F}_\theta} + (1-\alpha^0_{\theta'}) \underbrace{\left[\mu_\theta W^{RSW,S}_{\theta'} + (1-\mu_\theta) W^{RSW,F}_{\theta'}\right]}_{\geq \mu_\theta W^{RSW,S}_\theta + (1-\mu_\theta) W^{RSW,F}_\theta} \\
&\qquad \text{by } q^0\text{-feasibility of } W^0 \qquad\qquad \text{by } q\text{-feasibility of } W^{RSW}\ \forall q \in \Delta(\Theta) \\
&\quad -\alpha^0_\theta \left[\mu_\theta W^{0,S}_\theta + (1-\mu_\theta) W^{0,F}_\theta\right] - (1-\alpha^0_\theta)\left[\mu_\theta W^{RSW,S}_\theta + (1-\mu_\theta) W^{RSW,F}_\theta\right] \\
&\geq \left[\mu_\theta W^{0,S}_\theta + (1-\mu_\theta) W^{0,F}_\theta - \mu_\theta W^{RSW,S}_\theta - (1-\mu_\theta) W^{RSW,F}_\theta\right](\alpha^0_{\theta'} - \alpha^0_\theta) \geq 0,
\end{aligned}$$

where the first inequality holds because the wage scheme W^0 is q^0-feasible and W^{RSW} is feasible regardless of the agent's belief, and the second inequality holds because α_θ is type θ's optimal choice probability between W^0 and W^{RSW}.
Let $\tilde{v} \in \mathcal{V}(q^*)$ be the payoff implemented by $\tilde{W}$. By construction,

$$\tilde{v} = \alpha^0 \cdot v^0 + (1-\alpha^0) \cdot v^{RSW} \geq v^{RSW}.$$

However, this contradicts the conclusion of Lemma 4 that v^{RSW} is q^*-undominated. ■

Now we can prove the sufficient condition of Proposition 1. Fix any prior $p_0 \in \text{int}(\Delta(\Theta))$ and take any payoff vector $v \in \mathcal{V}^*(p_0)$. By definition, $v \geq v^{RSW}$. From Lemma 5, for any deviation $\tilde{\mathcal{C}}$, there exists a belief $\tilde{q} \in \Delta(\Theta)$ and a payoff $\tilde{v} \in \Gamma(\tilde{\mathcal{C}}, \tilde{q}) \subseteq \mathcal{V}(\tilde{q})$ such that $v^{RSW} \geq \tilde{v}$. Hence, for any off-path contract proposal, there is a belief that makes the deviation unprofitable for all types of the principal. ■

Proof of Proposition 2. Since $\mathcal{V}(p_0)$ is a convex polygon, so is $\mathcal{V}^*(p_0)$.[19] Therefore, I just

[19]Recall that $\mathcal{V}^*(p_0) = \{v \in \mathcal{V}(p_0) : v \geq v^{RSW}\}$.

need to characterize the extreme points, which is given by solving the following linear program

$$\max_{W\in[0,\bar{w}]^4} \omega_L\Big(\mu_L W_L^S + (1-\mu_L)W_L^F\Big) + \omega_H\Big(\mu_H W_H^S + (1-\mu_H)W_H^F\Big) \qquad \text{(Program 6-}\omega\text{)}$$

$$s.t. \sum_{\theta\in\Theta} p_0(\theta)\mu_\theta\Big(W_\theta^S - W_\theta^F\Big) \geq c, \qquad \text{(A-IC}_{p_0}\text{)}$$

$$\mu_\theta\Big(W_{\theta'}^S - W_\theta^S\Big) + (1-\mu_\theta)\Big(W_{\theta'}^F - W_\theta^F\Big) \geq 0,\ \forall\theta'\in\Theta, \qquad \text{(P-IC}_\theta,\ \forall\theta\text{)}$$

$$\mu_\theta\nu_S - \mu_\theta W_\theta^S - (1-\mu_\theta)W_\theta^F \geq v_\theta^{RSW}. \qquad \text{(PBE}_\theta,\ \forall\theta\text{)}$$

as the weights $\omega = (\omega_L, \omega_H)$ vary over $\mathbb{R}^2\backslash\{(0,0)\}$. Notice that the principal's payoff is now bounded below by v^{RSW} instead of the outside option.

For $\omega > 0$, it is clear that the maximum is attained when the PBE_θ constraints bind for all $\theta\in\Theta$. Hence, the solution is given by the RSW wages in Table 1. In fact, when $p_0(H) < p^*(H)$, this is the only solution for all weights $\omega \neq 0$. When $p_0(H) \geq p^*(H)$, the solution is summarized in Figure 4 below.

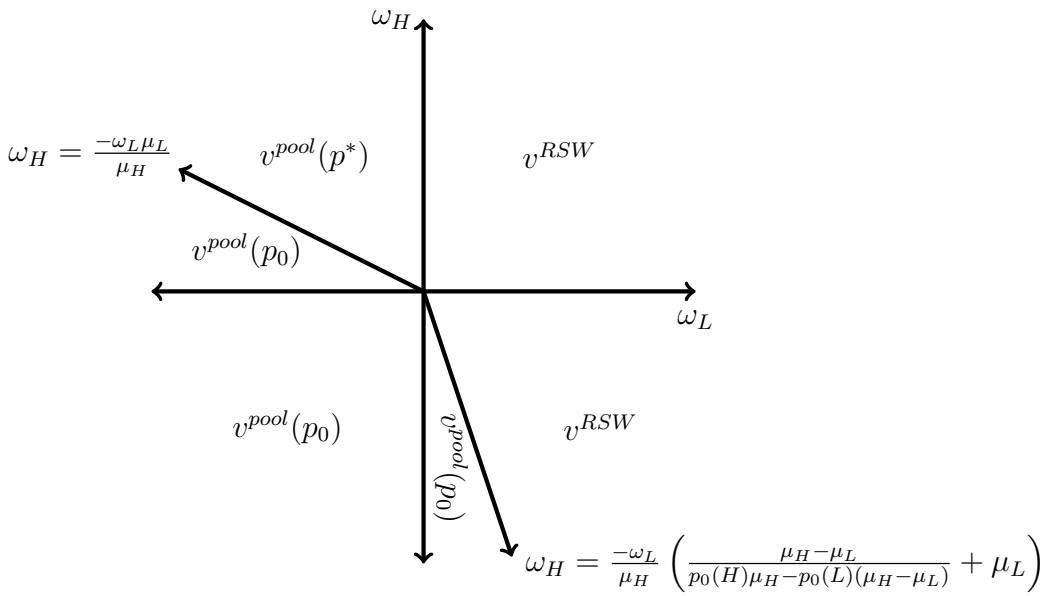


Figure 4: Solutions to (Program 6-ω) when $p_0(H) \geq p^*(H)$.

■

Proof of Lemma 2. Suppose there exists a payoff $v \in \mathcal{V}(p_0)$ that is implementable by an SNP mechanism. Then $v \geq v^{RSW}$. Otherwise, the RSW mechanism presented in Table 1 along

with a belief q satisfying

$$q(\theta) = \frac{p_0(\theta)}{\sum_{\theta' \in S(v^{RSW},v)} p_0(\theta')} \times \mathbb{1}_{S(v^{RSW},v)}(\theta)$$

would satisfy conditions (a)-(b) of Definition 3.[20] By Proposition 1, $v \in \mathcal{V}^*(p_0)$.
From Corollary 1, there is a unique PBE payoff, $\bar{v}^{p_0}$, that Pareto dominates all other payoffs in $\mathcal{V}^*(p_0)$. If $v \neq \bar{v}^{p_0}$, then the belief $q = p_0$ along with a deviation to $\bar{v}^{p_0}$ would satisfy conditions (a)-(b), contradicting the assumption that v is SNP implementable. Hence, $v = \bar{v}^{p_0}$. ■

Proof of Lemma 3. Fix some finite level of limited liability $\underline{w} \in (-\infty, 0]$ and a prior $p_0 \in \text{int}(\Delta(\Theta))$. Suppose there exists a first-best-surplus extracting equilibrium mechanism M. Then, all types must recommend action e_{l^*}. Furthermore, $M \triangleq \langle \boldsymbol{w}(\theta_t) \rangle_{t=1}^T$ satisfies

(a) $\boldsymbol{\mu}(\theta_t, e_{l^*}) \cdot \big(\boldsymbol{\nu} - \boldsymbol{w}(\theta_t)\big) = S_t^* = \boldsymbol{\mu}(\theta_t, e_{l^*}) \cdot \boldsymbol{\nu} - c_{l^*}, \ \ \forall t = 1, 2, \ldots, T,$

(b) $\boldsymbol{\mu}(\theta_t, e_{l^*}) \cdot \big(\boldsymbol{\nu} - \boldsymbol{w}(\theta_t)\big) \geq \boldsymbol{\mu}(\theta_t, e_{l^*}) \cdot \big(\boldsymbol{\nu} - \boldsymbol{w}(\theta_{\hat{t}})\big), \ \ \forall \hat{t} \neq t, \ \ \forall t = 1, 2, \ldots, T,$

(c) $\sum_{t=1}^T p_0(\theta_t)\big(\boldsymbol{\mu}(\theta_t, e_{l^*}) - \boldsymbol{\mu}(\theta_t, e_l)\big) \cdot \boldsymbol{w}(\theta_t) \geq c_{l^*} - c_l, \ \ \forall l \neq l^*,$ and

(d) $w^n(\theta_t) \geq \underline{w}, \ \ \forall t = 1, 2, \ldots, T, \ \ \forall n = 1, 2, \ldots, N,$

where (a) follows from each type extracting the first-best surplus, (b) follows from P-IC$_t^{\hat{t}}$, (c) follows from A-IC$_{p_0}^l$, and (d) follows the limited liability constraints. We get *i*-*iv* of Lemma 3 by setting $k_t^n = w^n(\theta_t) - c_{l^*}$ for $n = 1, 2, \ldots, N$ and $t = 1, 2, \ldots, T$.

To see sufficiency, note that if Lemma 3 holds, then there is a mechanism $M \triangleq \langle \boldsymbol{w}(\theta_t) \rangle_{t=1}^T$ that extracts the surplus and is p_0-feasible. From Wagner et al. (2015), a p_0-feasible surplus extracting mechanism is strongly neologism proof, and therefore, an equilibrium mechanism. ■

[20] Given some set $S \subseteq Y$ and some variable $y \in Y$, $\mathbb{1}_S(y) = 1$ if $y \in S$ and 0 otherwise