Economic Theory Bulletin (2021) 9:119–142
https://doi.org/10.1007/s40505-021-00201-3

RESEARCH ARTICLE

Informed principal, moral hazard, and limited liability

Teddy Mekonnen[1]

Received: 8 November 2019 / Accepted: 17 January 2021 / Published online: 9 February 2021

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.

Keywords Informed principal · Limited liability · Surplus extraction

JEL Classification D82 · D86

1 Introduction

In the standard model of moral hazard, informational asymmetries arise because the principal is unable to monitor the agent's effort. 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 have private information on how the agent's effort affects the distribution of observable outcomes, i.e., the production technology. For example, the principal may have private information about the difficulty of a task an agent is hired to accomplish. A principal who knows the task is easy may wish to reveal that information to encourage the agent to work, whereas a principal who knows the task is difficult may prefer not to disclose that information. Such opposing signaling incentives could lead to distortions. Nonetheless, Wagner et al. (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

✉ Teddy Mekonnen
mekonnen@brown.edu

[1] Department of Economics, Brown University, Providence, RI, USA

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's private information were made public or the agent had unlimited liability, there is an equilibrium in which the principal extracts the full surplus. In other words, the inability to extract the surplus is not solely due to the limited liability or the informed principal but the combination of both.

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. Under complete information, each type of the principal bears the cost of incentive provision alone. For some types, incentive provision is cheap, and for others, it is expensive. Under incomplete information with an informed principal, the types that found it expensive to provide incentives have an incentive to lie and mimic the types that can incentivize the agent cheaply. I show that in environments in which such incentives to lie arise, an informed principal cannot extract surplus for all full-support priors while respecting limited liability.

Essentially, some types have a comparative advantage in providing the work incentives to the agent while the other types have a comparative advantage in providing the truthfulness incentives to the principal. A type from the former group can pool with a type from the latter group and "trade" the cost of incentive provision. However, with limited liability, the types that have a comparative advantage in incentivizing the agent are constrained as punishments are not feasible, leading to a breakdown of trade and the impossibility result.

Several papers have noted the existence of specific equilibrium outcomes in which an informed principal fails to extract all the surplus. Karle et al. (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).

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

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 Success or Failure 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 $\nu_S > \nu_F = 0$.[1] 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:

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

[1] 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

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

Table 1 First-best mechanism under complete information

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

2.1 Complete 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 $v_\theta^{FB} = \mu_\theta v_S - c$. The equilibrium can be implemented by each type θ offering the agent an outcome contingent payment given by Table 1.

Remark 1 It is well understood that limited liability generally prevents full surplus extraction under complete information. Here, the principal is able to extract the surplus despite limited liability because a successful outcome ($x = S$) fully reveals the agent's choice to work. Hence, the failure of full-surplus extraction with an *informed* principal in the next section is not because of limitations by the production technology but because of how limited liability and signaling distortions interact.

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 $p_0 \in \operatorname{int} \Delta(\Theta)$.[2]

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 $\mathcal{C} \triangleq (\mathcal{S}, w)$ which specifies (i) a finite set of messages $\mathcal{S}$ the principal can send to the agent in the continuation game, and (ii) a payment rule $w : \mathcal{S} \times X \to \mathbb{R}_+$ from the principal to the agent at the end of the game which depends on the message she sent and the observed outcome.

The proposed contract itself can be informative of the principal's type since different types may propose different contracts. Upon observing the proposed contract $\mathcal{C}$, the agent updates his belief to a posterior $q^{\mathcal{C}} \in \Delta(\Theta)$.[3]

The proposed contract and the agent's posterior together then define a finite perfect-recall extensive-form continuation game $(\mathcal{C}, q^{\mathcal{C}})$ as described in Fig. 1: the principal first sends a message $s \in \mathcal{S}$. The agent, after observing the message, chooses his effort $e \in E$. An outcome $x \in X$, whose distribution depends on the principal's type and the agent's effort, is then realized. Finally, the principal pays the agent $w(s, x)$ according to the contract specifications.

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

[3] 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.

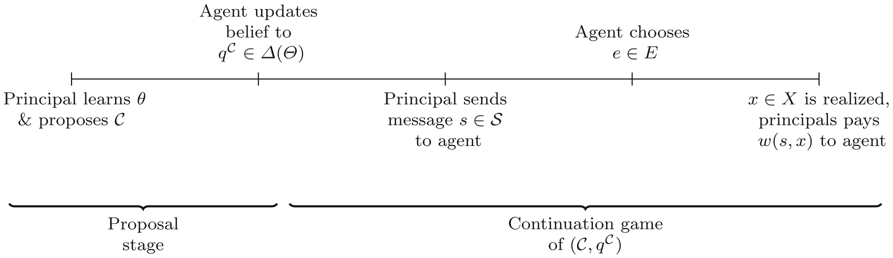
Agent updates belief to $q^{\mathcal{C}} \in \Delta(\Theta)$
Agent chooses $e \in E$
Principal learns θ & proposes $\mathcal{C}$
Principal sends message $s \in \mathcal{S}$ to agent
$x \in X$ is realized, principals pays $w(s, x)$ to agent
Proposal stage
Continuation game of $(\mathcal{C}, q^{\mathcal{C}})$

Fig. 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 $q^{\mathcal{C}} \in \Delta(\Theta)$ 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 $q^{\mathcal{C}}$ to the sequential equilibrium payoff set of the continuation game $(\mathcal{C}, q^{\mathcal{C}})$ is upper-hemicontinuous.[4] 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

$$w : \Theta \times E \times X \to \mathbb{R}_+$$

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 $w(\hat{\theta}, \bar{e}, x)$. Since the agent is risk-neutral, the payment scheme can be deterministic as long as we allow it to depend on the recommended action.

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

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

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

Similarly, if the principal reports her type truthfully, an agent with belief $q\in\Delta(\Theta)$ who gets a recommendation $\bar{e}\in \operatorname{supp}(r)$ and chooses effort $e\in E$ has an interim payoff given by

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

When the principal reports her type truthfully and the agent obeys the effort recommendation, I simplify the notation and write $V(\theta,M)$ and $U(\bar{e},q,M)$. 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 each recommendation the mechanism makes with positive probability, (ii) incentive compatible for the principal to report her type truthfully, and (iii) individually rational for each type of the principal:[5]

(i) $U(\bar{e},q,M)\geq U(e;\bar{e},q,M),\ \ \forall\bar{e}\in\operatorname{supp}(r),\forall e\in E.$
(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 $(\mathcal{C},q^{\mathcal{C}})$ is implementable by some $q^{\mathcal{C}}$-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.[6] 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 deviation 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 $\theta\in\Theta$.

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

[6] 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 recommend work ($\bar{e} = 1$) with probability one. 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,

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

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

$$\mathcal{M}(q) \triangleq \Big\{\langle w(\theta, x)\rangle_{\theta\in\Theta, x\in X} \in \mathbb{R}^4_+ \text{ satisfying A-IC}_q, \text{P-IC}_\theta, \text{P-IR}_\theta \forall\theta \in \Theta\Big\}.$$

Lemma 2 *$\mathcal{M}(q)$ is convex, and compact for all $q \in \Delta(\Theta)$, and $\cap_{q\in\Delta(\Theta)}\mathcal{M}(q) \neq \emptyset$.*

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.[7] Hence, it

[7] 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.

Table 2 RSW Mechanism

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

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 $\boldsymbol{v} \in \mathcal{V}(q)$ if, and only if, there exists a q-feasible mechanism M such that $\boldsymbol{v} = (v_L, v_H) = \Big(V(L, M), V(H, M)\Big)$. As $\mathcal{M}(q)$ is convex and compact, and as 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

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

A mechanism $M \in \bigcap_{q \in \Delta(\Theta)} \mathcal{M}(q)$ is feasible regardless of the agent's belief.[8] 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 Table2.

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, denoted by $\boldsymbol{v}^{RSW}$ such that

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

Let $\mathcal{V}^*(q) \triangleq \left\{\boldsymbol{v} \in \mathcal{V}(q) : \boldsymbol{v} \geq \boldsymbol{v}^{RSW}\right\}$ 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 2 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).

[8] I reformulate (2-θ) as a linear programming problem in the Appendix.

Proposition 1 *Given a prior $p_0 \in \operatorname{int} \Delta(\Theta)$, a payoff vector $\boldsymbol{v} \in \mathbb{R}^2$ is implementable by a PBE mechanism if, and only if, $\boldsymbol{v} \in \mathcal{V}^*(p_0)$. Furthermore, for any two beliefs $p_0, p_0' \in \Delta(\Theta)$ 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 $\boldsymbol{v} \in \mathcal{V}^*(p_0)$ is implementable by some p_0-feasible mechanism M, i.e., $V(\theta, M) = v_\theta$ for each $\theta \in \Theta$. 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, A-IC$_{p_0}$ is satisfied:

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

Using $v_L \geq v_L^{RSW}$ and Table 2,

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

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

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

establishing A-IC$_{p_0'}$. 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. Thus, M is p_0'-feasible and $\boldsymbol{v} \in \mathcal{V}^*(p_0')$. □

By definition, $\boldsymbol{v}^{RSW} \in \mathcal{V}^*(p_0)$ for any prior $p_0 \in \operatorname{int} \Delta(\Theta)$. Therefore, Proposition 1 implies that the RSW mechanism in Table 2 is always a PBE 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 $q \in \Delta(\Theta)$, let $\boldsymbol{v}^{pool}(q) = \big(v_L^{pool}(q), v_H^{pool}(q)\big) \in \mathcal{V}(q)$ be the payoff vector given by

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

Table 3 Pooling Mechanism

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

It can be implemented by the pooling mechanism $M^{pool}(q) \in \mathcal{M}(q)$ with wages given by Table 3.

It is straightforward to check that for any two beliefs $q, q' \in \Delta(\Theta)$ with $q'(H) > q(H)$, $\boldsymbol{v}^{pool}(q') > \boldsymbol{v}^{pool}(q)$.

Proposition 2 *There is a unique belief $p^* \in \operatorname{int} \Delta(\Theta)$ with $p^*(H) = \frac{\mu_H - \mu_L}{2\mu_H - \mu_L}$ such that for any prior $p_0 \in \operatorname{int} \Delta(\Theta)$, the equilibrium payoff set $\mathcal{V}^*(p_0)$ is given by*

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

where $\operatorname{conv}(\cdot)$ *is the convex hull.*

An immediate consequence of Proposition 2 is that there is a unique PBE payoff vector (namely, $\boldsymbol{v}^{RSW}$) if, and only if, $p_0(H) < p^*(H)$. Figure 2a–c below represent how the equilibrium payoff set $\mathcal{V}^*(p_0)$ changes as a function of the agent's prior.

Corollary 1 *For any prior $p_0 \in \operatorname{int} \Delta(\Theta)$, there is a unique Pareto-dominant PBE payoff vector $\bar{\boldsymbol{v}}^{\boldsymbol{p_0}}$, i.e., $\bar{\boldsymbol{v}}^{\boldsymbol{p_0}} \geq \boldsymbol{v}$ for all $\boldsymbol{v} \in \mathcal{V}^*(p_0)$.*

Proof **Case 1:** $p_0(H) < p^*(H)$.
$\boldsymbol{v}^{RSW}$ is the unique PBE payoff. The result is immediate by setting $\bar{\boldsymbol{v}}^{\boldsymbol{p_0}} = \boldsymbol{v}^{RSW}$.
Case 2: $p_0(H) = p^*(H)$.
From Proposition 2, $\boldsymbol{v} \in \mathcal{V}^*(p_0) = \operatorname{conv}\Big(\big\{\boldsymbol{v}^{RSW}, \boldsymbol{v}^{pool}(p^*)\big\} \Big)$. Note that $v_L^{RSW} < v_L^{pool}(p^*)$ and $v_H^{RSW} = v_H^{pool}(p^*)$. Thus, $\boldsymbol{v}^{pool}(p^*) \gneq \boldsymbol{v}$ for all $\boldsymbol{v} \in \mathcal{V}^*(p^*) \backslash \{\boldsymbol{v}^{pool}(p^*)\}$. We get the result by setting $\bar{\boldsymbol{v}}^{\boldsymbol{p_0}} = \boldsymbol{v}^{pool}(p^*)$.
Case 3: $p_0(H) > p^*(H)$.
In this case, $\boldsymbol{v}^{pool}(p_0) > \boldsymbol{v}^{pool}(p^*) \geq \boldsymbol{v}^{RSW}$. Thus, $\bar{\boldsymbol{v}}^{\boldsymbol{p_0}} = \boldsymbol{v}^{pool}(p_0) > \boldsymbol{v}$ for all $\boldsymbol{v} \in \mathcal{V}^*(p_0) \backslash \{\bar{\boldsymbol{v}}^{\boldsymbol{p_0}}\}$. □

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

To gain some intuition, recall the full information game in Sect. 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

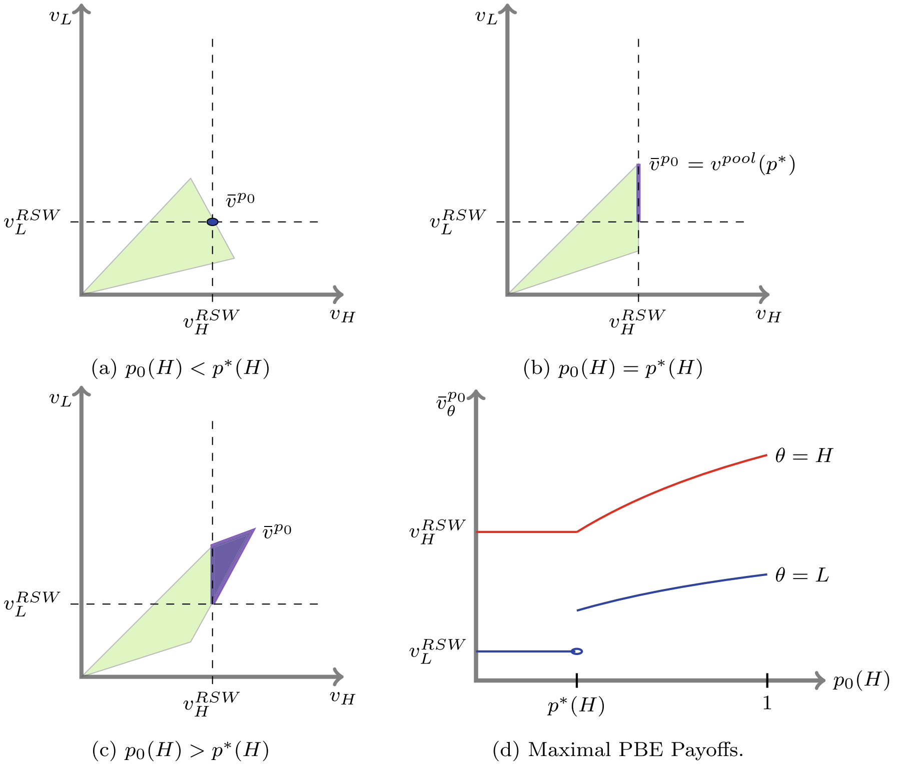

Fig. 2 The green area represents $\mathcal{V}(p_0)$, payoffs implementable by p_0-feasible mechanisms. The blue area represents $\mathcal{V}^*(p_0)$, the subset of payoffs implementable by PBE mechanisms

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 is too pessimistic, pooling is too costly and the high type prefers the separating mechanism in Table 2. Otherwise, the pooling mechanism in Table 3 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 $k \in \mathbb{R}^2$ and a type θ^* such that

$$\begin{aligned}
(i)&\ \ (1-\mu_{\theta^*}e)k_1 + \mu_{\theta^*}ek_2 = 0 \text{ for } e = 1,\\
(ii)&\ \ (1-\mu_{\theta^*}e)k_1 + \mu_{\theta^*}ek_2 < 0 \text{ for } e = 0, \text{ and}\\
(iii)&\ \ (1-\mu_{\theta}e)k_1 + \mu_{\theta}ek_2 \geq 0 \text{ for } e = 1 \text{ and } \theta \neq \theta^*.
\end{aligned}$$

Table 4 Full surplus extraction under unlimited liability

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

The current model satisfies these conditions with $k_1 = -\mu_L$ and $k_2 = 1 - \mu_L$ and $\theta^* = \theta_L$. The principal can implement $\boldsymbol{v}^{FB}$ through the mechanism $M^{WMT} = \langle w(\theta, x)\rangle_{\theta\in\Theta, x\in X}$ with Table 4.[9]

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

Proof From Proposition 2, $\mathcal{V}^*(p_0) = \{\boldsymbol{v}^{RSW}\}$ when $p_0(H) < p^*(H)$, and $\boldsymbol{v}^{RSW} \neq \boldsymbol{v}^{FB}$. From Corollary 1, we have $\bar{\boldsymbol{v}}^{\boldsymbol{p_0}} = \boldsymbol{v}^{pool}(p_0) \geq \boldsymbol{v}$ for all $\boldsymbol{v} \in \mathcal{V}^*(p_0)$ 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 is θ. In either case, $\boldsymbol{v}^{FB} \notin \mathcal{V}^*(p_0)$. □

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 4) 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 Generalization

In this section, I consider a T-types×L-actions×N-outcomes model and show a general impossibility result that extends Proposition 3. Let $\Theta \triangleq \{\theta_1, \theta_2, \ldots, \theta_T\}$ be the different types of the principal, and let $p_0 \in \operatorname{int} \Delta(\Theta)$ denote the agent's prior. Let $E \triangleq \{e_1, e_2, \ldots, e_L\}$ be the agent's effort space with associated costs

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

$0 = c(e_1) \leq c(e_2) \leq \ldots \leq c(e_L)$. Finally, let $\boldsymbol{v} \in \mathbb{R}^N$ be the vector of revenues with v_n the revenue associated with outcome $n = 1, 2, \ldots, N$.

Given the principal's type $\theta \in \Theta$ and the agent's effort $e \in E$, let $\boldsymbol{\mu}(\theta, e) = \big(\mu^n(\theta, e)\big)_{n=1}^N$ be the probability distribution over outcomes. For a wage vector $\boldsymbol{w} \in \mathbb{R}^N$, the principal's payoff is given by $\boldsymbol{\mu}(\theta, e) \cdot (\boldsymbol{v} - \boldsymbol{w})$ while the agent's payoff is given by $\boldsymbol{\mu}(\theta, e) \cdot \boldsymbol{w} - c(e)$. I assume the principal faces some finite level of limited liability $\underline{w} \in (\infty, 0]$ such that wages must satisfy $w \geq \underline{w}$. The outside option for the agent is denoted by $\underline{u} \geq 0$ and the outside option for the principal is normalized to 0.

Let $S(\theta, e) = \boldsymbol{\mu}(\theta, e) \cdot \boldsymbol{v} - c(e) - \underline{u}$ be the surplus generated between a principal of type $\theta \in \Theta$ and an agent who chooses action $e \in E$. For each $\theta \in \Theta$, let

$$e^*(\theta) = \arg\max_{e \in E} S(\theta, e).$$

To simplify exposition, I assume that $e^*(\theta)$ is unique; allowing for multiple efficient actions for a type would not change the results but would require random recommendations. Additionally, I assume that $S(\theta, e^*(\theta)) \geq 0$ for all $\theta \in \Theta$. In other words, it is efficient for all types to hire the agent, which generalizes (1) to the case where different types have different efficient actions. Let $E^* = \cup_{\theta \in \Theta} e^*(\theta)$ be the subset of efficient actions and let $\tilde{\Theta}(e) = \{\theta \in \Theta : e^*(\theta) = e\}$ be the subset of types whose efficient action is $e \in E$. Clearly, $\tilde{\Theta}(e) \neq \emptyset$for $e \in E^*$.

As the interest of the paper is in the first-best surplus extraction, I will focus on direct revelation mechanisms M in which each type $\theta \in \Theta$ recommends effort $e^*(\theta)$ and offers the wage vector $\boldsymbol{w}(\theta) \in \mathbb{R}^N$ such that $w_n(\theta) \geq \underline{w}$ for $n = 1, 2, \ldots, N$. If the agent obeys the effort recommendation, the principal's payoff from reporting $\hat{\theta}$ when her true type is θ is given by

$$V(\hat{\theta}; \theta, M) = \boldsymbol{\mu}(\theta, e^*(\hat{\theta})) \cdot \big(\boldsymbol{v} - \boldsymbol{w}(\hat{\theta})\big).$$

Similarly, if the principal reports her type truthfully, an agent with belief $q \in \Delta(\Theta)$ who gets a recommendation $\bar{e} \in E^*$ and chooses effort $e \in E$ has an interim payoff given by

$$U(e; \bar{e}, q, M) = \sum_{\theta \in \Theta} q(\theta|\bar{e}) \boldsymbol{\mu}(\theta, e) \cdot \boldsymbol{w}(\theta) - c(e)$$

where

$$q(\theta|\bar{e}) = \begin{cases} \frac{q(\theta)}{\sum_{\theta' \in \tilde{\Theta}(\bar{e})} q(\theta')} & \text{if } \theta \in \tilde{\Theta}(\bar{e}) \\ 0 & \text{otherwise} \end{cases}.$$

When the principal reports her type truthfully and the agent obeys the effort recommendation, I simplify the notation and write $V(\theta, M)$ and $U(\bar{e}, q, M)$. The mechanism M is p_0-feasible for prior $p_0 \in \operatorname{int} \Delta(\Theta)$ if:

1. the agent is willing to obey each recommendation of the mechanism, i.e.,

$$U(\bar{e}, p_0, M) \geq U(e; \bar{e}, p_0, M), \ \ \forall \bar{e} \in E^*, \forall e \in E \tag{A-IC$^e_{\bar{e}}$}$$

2. the agent is willing to participate in the mechanism ex-ante, i.e.,

$$\sum_{\theta \in \Theta} p_0(\theta) U(e^*(\theta), p_0, M) \geq \underline{u} \tag{A-IR}$$

3. each type of the principal is willing to report truthfully, i.e.,

$$V(\theta, M) \geq V(\hat{\theta}; \theta, M), \ \ \forall \theta, \hat{\theta} \in \Theta \tag{P-IC$^{\hat{\theta}}_{\theta}$}$$

4. and each type of the principal is wiling to participate in the mechanism, i.e.,

$$V(\theta, M) \geq 0, \ \ \forall \theta \in \Theta. \tag{P-IR$_\theta$}$$

In contrast to Sect. 3, I do not seek to characterize the entire equilibrium payoff set in the $T \times L \times N$ model. Instead, in Lemma 3, I provide a necessary and sufficient condition under which first-best surplus extraction is possible by all types of the principal. The conditions of Lemma 3 apply for any level of liability including unlimited liability.[10] Hence, these conditions subsume the sufficient full rank conditions in Wagner et al. (2015). Under an additional assumption, I then show that these necessary conditions cannot be satisfied for all full-support prior beliefs.

Lemma 3 *There exists a PBE mechanism M with $V(\theta, M) = S(\theta, e^*(\theta))$ for each $\theta \in \Theta$ if, and only if, there exist vectors $\langle \boldsymbol{k}_{\theta_1}, \boldsymbol{k}_{\theta_2}, \ldots, \boldsymbol{k}_{\theta_T} \rangle$ with $\boldsymbol{k}_\theta \in \mathbb{R}^N$ such that for all $\theta \in \Theta$,*

(i) $\boldsymbol{\mu}(\theta, e^*(\theta)) \cdot \boldsymbol{k}_\theta = 0$,
(ii) $\boldsymbol{\mu}(\hat{\theta}, e^*(\theta)) \cdot \boldsymbol{k}_\theta \geq S(\hat{\theta}, e^*(\theta)) - S(\hat{\theta}, e^*(\hat{\theta}))$, *for all* $\hat{\theta} \in \Theta$,
(iii) $\sum_{\theta' \in \Theta} p_0(\theta' | e^*(\theta)) \boldsymbol{\mu}(\theta', e) \cdot \boldsymbol{k}_{\theta'} \leq c(e) - c(e^*(\theta))$ *for all* $e \in E$, *and*
(iv) $k^n_\theta \geq \underline{w} - \underline{u} - c(e^*(\theta))$ *for all* $n = 1, 2, \ldots, N$.

In the $2 \times 2 \times 2$ model, the high type was able to get her first-best payoff under full information using the mechanism in Table 1. However, the mechanism could not be implemented under incomplete information because the low type would get more than her first-best payoff by mimicking the high type. I extend this idea to the $T \times L \times N$ model: there exists a "preferred" type θ^* such that if a mechanism allows this type to get her first-best payoff under full information while incentivizing the agent to choose a costly but efficient effort $e^*(\theta^*)$ instead of a costless effort e_1 (i.e., "working v.s. shirking"), then there exists another type θ' that would strictly prefer to mimic θ^*.

Assumption 1 There exists a type θ^* such that $c(e^*(\theta^*)) > 0$, and for any $\boldsymbol{k} \in \mathbb{R}^N$ satisfying

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

(i) $\boldsymbol{\mu}(\theta, e^*(\theta)) \cdot \boldsymbol{k} = 0$, and
(ii) $\boldsymbol{\mu}(\theta, e_1) \cdot \boldsymbol{k} \leq c(e_1) - c(e^*(\theta)) = -c(e^*(\theta))$,

there exist a type $\theta' \in \Theta$ such that $\boldsymbol{\mu}(\theta', e^*(\theta^*)) \cdot \boldsymbol{k} < S(\theta', e^*(\theta^*)) - S(\theta', e^*(\theta'))$.

In other words, Assumption 1 implies that it is impossible for all types of the principal to independently incentivize the agent to work and to incentivize other types to be honest. Intuitively, since the distribution of outcomes is type-dependent, the cost of incentivizing the agent to "work" instead of "shirk" is not uniform across types. For some types, providing incentives is cheap while for others it is expensive. Under incomplete-information, the types that find it expensive to incentivize the agent will mimic the types that find it cheap to incentivize the agent. Thus, the types have to "trade" the cost of satisfying the agent's incentive constraint with the cost of satisfying the principal's incentive constraint.

Example 1 Assumption 1 is a fairly strong condition but it is nonetheless satisfied in a familiar class of moral-hazard problems introduced by Grossman and Hart (1983). Let $\boldsymbol{\mu}(\theta, e) = \lambda(\theta, e)\bar{\boldsymbol{\mu}} + \big(1 - \lambda(\theta, e)\big)\underline{\boldsymbol{\mu}}$ where $\bar{\boldsymbol{\mu}} = (\bar{\mu}_n)_{n=1}^N$ and $\underline{\boldsymbol{\mu}} = (\underline{\mu}_n)_{n=1}^N$ are two distributions of outcomes, and $\lambda : \Theta \times E \to [0, 1]$. Assumption 1 is satisfied if $\lambda(\theta, e)$ is monotone in both θ and e, and there exist $\theta', \theta'' \in \Theta$ such that $e^*(\theta') = e^*(\theta'')$ with $c(e^*(\theta')) > 0$. The $2 \times 2 \times 2$ model can be seen as a special case of this setting. I provide details in the Appendix.

It is possible for both the full rank conditions in Wagner et al. (2015) and Assumption 1 to hold simultaneously as in the $2 \times 2 \times 2$ model: the low type satisfies the full-rank conditions as discussed in Sect. 3 while the high type satisfies the conditions of Assumption 1. This is why within the same model, full-surplus extraction could be possible with unlimited liability but impossible under limited liability.

Proposition 4 *Suppose Assumption 1 holds. There exists a non-empty set of beliefs*

$$\mathcal{P}(\underline{w}, \underline{u}) = \left\{ q \in \operatorname{int} \Delta(\Theta) : q(\theta^*) \geq \frac{\underline{u} - \underline{w}}{\underline{u} - \underline{w} + c(e^*(\theta^*))} \right\}$$

such that for any full-support prior $p_0 \in \mathcal{P}(\underline{w}, \underline{u})$, there is no PBE mechanism M with $V(\theta, M) = S(\theta, e^*(\theta))$ for all $\theta \in \Theta$.

Proof Fix some prior $p_0 \in \mathcal{P}(\underline{w}, \underline{u})$. Suppose there exists a PBE mechanism M with $V(\theta, M) = S(\theta, e^*(\theta))$ for each $\theta \in \Theta$. By Lemma 3-(iii), there necessarily exist vectors $\langle \boldsymbol{k}_{\theta_1}, \boldsymbol{k}_{\theta_2}, \ldots, \boldsymbol{k}_{\theta_T} \rangle$ such that

$$\sum_{\theta' \in \Theta} p_0\big(\theta' | e^*(\theta)\big) \boldsymbol{\mu}(\theta', e_1) \cdot \boldsymbol{k}_{\theta'} \leq -c\big(e^*(\theta)\big)$$

for each $\theta \in \Theta$. If $\boldsymbol{\mu}(\theta, e_1) \cdot \boldsymbol{k}_\theta \leq -c\big(e^*(\theta)\big)$ for each $\theta \in \Theta$, then by Assumption 1, there would be a preferred type $\theta^* \in \Theta$ and another type $\theta' \in \Theta$ such that

$$\boldsymbol{\mu}(\theta', e^*(\theta^*)) \cdot \boldsymbol{k}_{\theta^*} < S(\theta', e^*(\theta^*)) - S(\theta', e^*(\theta')),$$

which would violate Lemma 3-(ii). Hence, $\boldsymbol{\mu}(\theta^*, e_1) \cdot \boldsymbol{k}_{\theta^*} > -c\big(e^*(\theta^*)\big)$. Since Lemma 3-$(iii)$ must hold, we necessarily have

$$\sum_{\theta' \in \Theta \backslash \{\theta^*\}} p_0\big(\theta' | e^*(\theta^*)\big) \boldsymbol{\mu}(\theta', e_1) \cdot \boldsymbol{k}_{\theta'} < -c\big(e^*(\theta^*)\big). \tag{4}$$

If $|\tilde{\Theta}\big(e^*(\theta^*)\big)| = 1$, then $p_0(\theta'|e^*(\theta^*)) = 0$ for all $\theta' \in \Theta \backslash \{\theta^*\}$ by Bayes-rule and (4) would be violated. Thus, $|\tilde{\Theta}\big(e^*(\theta^*)\big)| > 1$ and $p_0(\theta^*|e^*(\theta^*)) < 1$. Additionally, Lemma 3-$(iv)$ implies that

$$\sum_{\theta' \in \Theta \backslash \{\theta^*\}} p_0\big(\theta' | e^*(\theta^*)\big) \boldsymbol{\mu}(\theta', e_1) \cdot \boldsymbol{k}_{\theta'} \geq \sum_{\theta' \in \Theta \backslash \{\theta^*\}} p_0\big(\theta' | e^*(\theta^*)\big) \big(\underline{w} - \underline{u} - c(e^*(\theta'))\big).$$

Putting together with (4), we have

$$\begin{aligned} &\sum_{\theta' \in \Theta \backslash \{\theta^*\}} p_0\big(\theta' | e^*(\theta^*)\big) \big(\underline{w} - \underline{u} - c(e^*(\theta'))\big) < -c\big(e^*(\theta^*)\big) \\ \implies &\big(\underline{w} - \underline{u} - c(e^*(\theta^*))\big) \sum_{\theta' \in \Theta \backslash \{\theta^*\}} q\big(\theta' | e^*(\theta)\big) < -c\big(e^*(\theta^*)\big) \end{aligned}$$

where the second line follows because $p_0(\theta'|e^*(\theta^*)) > 0$ only if $\theta' \in \tilde{\Theta}(e^*(\theta^*))$, i.e., $e^*(\theta') = e^*(\theta^*)$. However, notice that the last line implies

$$\begin{aligned} &\Big(\underline{w} - \underline{u} - c(e^*(\theta^*))\Big)\Big(1 - p_0(\theta^*|e^*(\theta^*)\Big) < -c(e^*(\theta^*)) \\ &\qquad \implies p_0(\theta^*|e^*(\theta^*)) < \frac{\underline{u} - \underline{w}}{\underline{u} - \underline{w} + c(e^*(\theta^*)} \end{aligned}$$

which is impossible as $p_0 \in \mathcal{P}(\underline{w}, \underline{u})$. Hence, there exists no vector $\langle \boldsymbol{k}_{\theta_1}, \boldsymbol{k}_{\theta_2}, \ldots, \boldsymbol{k}_{\theta_T} \rangle$ that satisfies Lemma 3, which in turn implies there is no PBE mechanism that allows each type of the principal to get her first-best payoff. □

Notice that when $\underline{u} = \underline{w} = 0$, then full-surplus extraction is impossible for any full-support prior. As the limited liability constraint becomes less binding, i.e., $\underline{w}$ decreases, the set of beliefs for which full-surplus extraction is impossible shrinks.

Acknowledgements I am grateful to Eddie Dekel, Asher Wolinsky, and Bruno Strulovici for their invaluable comments.

5 Appendix

Proof of Lemma 1 Take any q-feasible mechanism $M \triangleq (r, w)$ and let $\boldsymbol{v} = \Big(V(L, M), V(H, M)\Big)$. 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

$$\begin{aligned} &U(1, q, \tilde{M}) - U(0; 1, q, \tilde{M}) \\ &= \sum_{\theta\in\Theta} q(\theta)\Big\{\mu_\theta\Big(\tilde{w}(\theta, 1, S) - \tilde{w}(\theta, 1, F)\Big) - c\Big\} \\ &= \underbrace{\sum_{\theta\in\Theta} q(\theta)r(1|\theta)\Big\{\mu_\theta\Big(w(\theta, 1, S) - w(\theta, 1, F)\Big) - c\Big\}}_{\substack{\geq 0 \\ \text{by } q\text{-feasibility of } M}} \\ &\quad + \sum_{\theta\in\Theta} q(\theta)r(0|\theta)\underbrace{\Big\{\mu_\theta \nu_S - c\Big\}}_{\substack{\geq 0 \\ \text{by (1)}}} \geq 0. \end{aligned}$$

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

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

and truthful reporting remains incentive compatible for all types of the principal under $\tilde{M}$. Thus, $\tilde{M}$ is q-feasible and implements the payoff $\boldsymbol{v} = \Big(V(L, M), V(H, M)\Big)$. □

Proof of Lemma 2 To show that $\cap_{q\in\Delta(\Theta)}\mathcal{M}(q) \neq \varnothing$, note that 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 A-IC$_q$, P-IC$_\theta$ and P-IR$_\theta$ constraints for all $q \in \Delta(\Theta)$ and all $\theta \in \Theta$.

Let $\bar{w} = \max\Big\{\nu_s, \frac{\mu_H}{1-\mu_H}\nu_s\Big\}$. Note that for all $\theta \in \Theta$ and all $x \in X$, $w(\theta, x) \geq 0$ by limited liability and $w(\theta, x) \leq \bar{w}$ by P-IR$_\theta$ for each θ. Hence, $\mathcal{M}(q)$ is a non-empty intersection of closed half-spaces with $\mathcal{M}(q) \subseteq [0, \bar{w}]^4$. Thus, it is convex and compact. □

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

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

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

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

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

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:

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

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

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

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 2 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 $q^* \in \operatorname{int} \Delta(\Theta)$ such that $\boldsymbol{v}^{RSW}$ is q^*-undominated, i.e., there exists no payoff $\boldsymbol{v} \in \mathcal{V}(q^*)$ 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

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

$$s.t. \ \mu_\theta \Big(W_\theta^S - W_\theta^F \Big) \geq c,$$

$$(\text{A-IC}_{\delta_\theta}, \forall \theta \in \Theta)$$

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

$$(\text{P-IC}_\theta, \forall \theta \in \Theta)$$

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 2, 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

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

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

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

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 $\boldsymbol{v}^{RSW}$ is q^*-undominated. □

Let $\mathcal{V} \triangleq \text{cl}\left(\text{conv}\Big(\bigcup_{q \in \Delta(\Theta)} \mathcal{V}(q)\Big)\right)$ be the convex closure of all feasible payoffs. For any payoff vector $\boldsymbol{v} \in \mathcal{V}$, there is a (random) direct revelation mechanism that can implement $\boldsymbol{v}$.[11] 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

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

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

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

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

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

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

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

The correspondence $T^\epsilon_{\mathcal{C}}$ is upper-hemicontinuous, convex-valued, and closed. Therefore, it has a fixed point, $(\boldsymbol{v}^\epsilon, \boldsymbol{\alpha}^\epsilon, q^\epsilon)$.

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 $\boldsymbol{v}^1 \in \Gamma(\mathcal{C}, q^1)$. Pick a choice probability $\boldsymbol{\alpha}^1 \in \mathcal{A}^\epsilon(\boldsymbol{v}^1)$ which describes the principal's optimal behavior when choosing between $\boldsymbol{v}^1$ and $\boldsymbol{v}^{RSW}$. Based on this behavior, the agent updates his belief from q^* to $q^2 = \mathcal{Q}^\epsilon(\boldsymbol{\alpha}^1) \in \Delta(\Theta)$. We now have a different continuation game $(\mathcal{C}, q^2)$. Pick a new sequential equilibrium payoff vector $\boldsymbol{v}^2 \in \Gamma(\mathcal{C}, q^2)$ and a new probability $\boldsymbol{\alpha}^2 \in \mathcal{A}^\epsilon(\boldsymbol{v}^2)$ of choosing $\boldsymbol{v}^2$ over $\boldsymbol{v}^{RSW}$. Based on this behavior, the agent updates his belief from q^* to $q^3 = \mathcal{Q}^\epsilon(\boldsymbol{\alpha}^2) \in \Delta(\Theta)$. And so on.

The fixed point $(\boldsymbol{v}^\epsilon, \boldsymbol{\alpha}^\epsilon, q^\epsilon)$ 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 $q^\epsilon = \mathcal{Q}^\epsilon(\boldsymbol{\alpha}^\epsilon)$. This results in the continuation game $(\mathcal{C}, q^\epsilon)$ and the subsequent sequential equilibrium outcome of the continuation game $\boldsymbol{v}^\epsilon \in \Gamma(\mathcal{C}, q^\epsilon)$. Based on this outcome, the constrained probability of deviating from the RSW mechanism to $\mathcal{C}$ is optimal, i.e., $\boldsymbol{\alpha}^\epsilon \in \mathcal{A}^\epsilon(\boldsymbol{v})$.

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

Proof Suppose not! Then, there exists a contract $\mathcal{C}$ such that for all beliefs $q \in \Delta(\Theta)$ and all payoffs $\boldsymbol{v} \in \Gamma(\mathcal{C}, q)$, some type of the principal strictly prefers $\boldsymbol{v}$ to $\boldsymbol{v}^{RSW}$.

Construct a fixed point $(\boldsymbol{v}^\epsilon, \boldsymbol{\alpha}^\epsilon, q^\epsilon)$ for $\epsilon > 0$ as described above and let

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

The limit is well defined: $\forall q \in \Delta(\Theta)$ and $\forall \boldsymbol{v} \in \Gamma(\mathcal{C}, q)$, some type strictly prefers $\boldsymbol{v}$ to $\boldsymbol{v}^{RSW}$ implies

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

Thus, q^0 is a well-defined probability distribution over Θ . As $\boldsymbol{v}^0 \in \Gamma(\mathcal{C}, q^0) \subseteq \mathcal{V}(q^0)$, 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:

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

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

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

where the first inequality holds because α_θ is type θ's optimal choice probability between W^0 and W^{RSW} and the last inequality holds because the wage scheme W^0 is q^0-feasible and W^{RSW} is feasible regardless of the agent's belief.

Let $\tilde{\boldsymbol{v}} \in \mathcal{V}(q^*)$ be the payoff implemented by $\tilde{W}$. By construction,

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

However, this contradicts the conclusion of Lemma 4 that $\boldsymbol{v}^{RSW}$ 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 $\boldsymbol{v} \in \mathcal{V}^*(p_0)$. By definition, $\boldsymbol{v} \geq \boldsymbol{v}^{RSW}$. From Lemma 5, for any deviation $\tilde{\mathcal{C}}$, there exists a belief $\tilde{q} \in \Delta(\Theta)$ and a payoff $\tilde{\boldsymbol{v}} \in \Gamma(\tilde{\mathcal{C}}, \tilde{q}) \subseteq \mathcal{V}(\tilde{q})$ such that $\boldsymbol{v}^{RSW} \geq \tilde{\boldsymbol{v}}$. 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 $\mathcal{V}(p_0)$ is non-empty, convex, and compact by Lemma 2. Since $\mathcal{V}^*(p_0) = \{\boldsymbol{v} \in \mathcal{V}(p_0) : \boldsymbol{v} \geq \boldsymbol{v}^{RSW}\}$ and $\boldsymbol{v}^{RSW} \in \mathcal{V}(p_0)$, $\mathcal{V}^*(p_0)$ is also non-empty, convex, and compact. Hence, $\mathcal{V}^*(p_0)$ can be characterized by first finding its boundary and then taking its convex hull.

By the Supporting Hyperplane Theorem, any point on the boundary of $\mathcal{V}^*(p_0)$ is contained in its supporting hyperplane. In order to find the supporting hyperplane, we need the support function. For weights $\boldsymbol{\omega} = (\omega_L, \omega_H) \in \mathbb{R}^2 \backslash \mathbf{0}$, define

$$h_{\mathcal{V}^*(p_0)}(\boldsymbol{\omega}) = \max_{\boldsymbol{v} \in \mathcal{V}^*(p_0)} \boldsymbol{\omega} \cdot \boldsymbol{v} = \omega_L v_L + \omega_h v_h$$

as the support function of $\mathcal{V}^*(p_0)$. By definition, $v \in \mathcal{V}^*(p_0)$ if and only if there are wages $W \triangleq \langle W_\theta^x \rangle_{\theta \in \Theta, x \in X}$ such that

(a) $W \in \mathcal{M}(p_0)$, and
(a) $v_\theta = \mu_\theta v - \mu_\theta W_\theta^S - (1-\mu_\theta) W_\theta^F \geq v_\theta^{RSW}$ for each $\theta \in \Theta$.

Hence, for weights $\boldsymbol{\omega} = (\omega_L, \omega_H) \in \mathbb{R}^2 \backslash \mathbf{0}$, the support function is equivalent to solving the following linear program:

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

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

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

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

For $\boldsymbol{\omega} > 0$, it is clear that the maximum is attained when the PBE$_\theta$ constraints bind for all $\theta \in \Theta$. Hence, the solution is given by the RSW wages in Table 2. In fact, when $p_0(H) < p^*(H)$, this is the only solution for all weights $\boldsymbol{\omega} \neq \mathbf{0}$. When $p_0(H) \geq p^*(H)$, the solution is summarized in Fig. 3 below. □

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, each type $\theta \in \Theta$ must recommend action $e^*(\theta)$. Furthermore, the wages paid $\langle \boldsymbol{w}(\theta_t) \rangle_{t=1}^T$ must satisfy

(a) $\boldsymbol{\mu}(\theta, e^*(\theta)) \cdot \big(\boldsymbol{v} - \boldsymbol{w}(\theta)\big) = S(\theta, e^*(\theta)) = \boldsymbol{\mu}(\theta), e^*(\theta)) \cdot \boldsymbol{v} - c(e^*(\theta)) - \underline{u}$, $\forall \theta \in \Theta$,

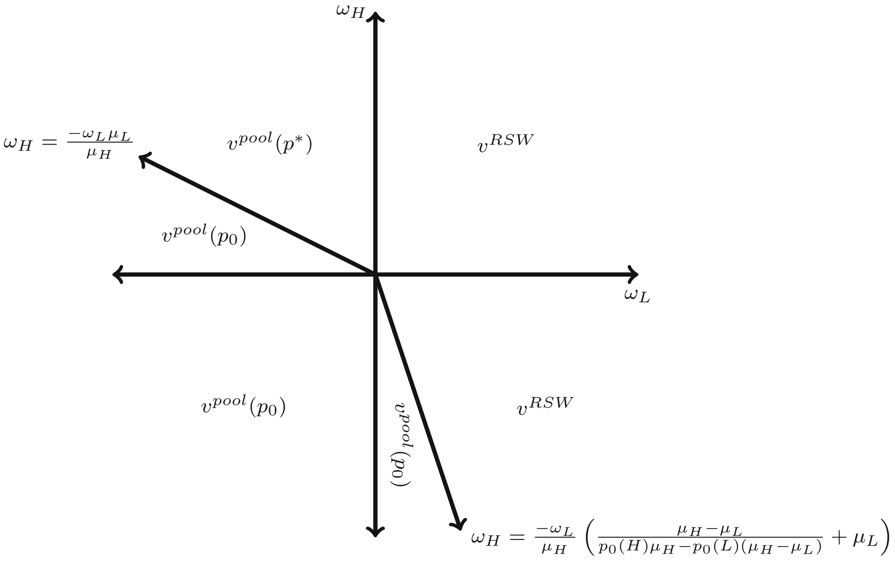
ω_H
$\omega_H = \frac{-\omega_L \mu_L}{\mu_H}$
$v^{pool}(p^*)$
v^{RSW}
$v^{pool}(p_0)$
ω_L
$v^{pool}(p_0)$
$v^{pool}(p_0)$
v^{RSW}
$\omega_H = \frac{-\omega_L}{\mu_H}\left(\frac{\mu_H-\mu_L}{p_0(H)\mu_H - p_0(L)(\mu_H-\mu_L)} + \mu_L\right)$

Fig. 3 Solutions to (Program 6-ω) when $p_0(H) \geq p^*(H)$

(b) $\boldsymbol{\mu}(\hat{\theta}, e^*(\hat{\theta})) \cdot \big(\boldsymbol{v} - \boldsymbol{w}(\hat{\theta})\big) \geq \boldsymbol{\mu}(\hat{\theta}, e^*(\theta)) \cdot \big(\boldsymbol{v} - \boldsymbol{w}(\theta)\big)$ $\forall \hat{\theta}, \theta \in \Theta$,
(c) $\sum_{t=1}^{T} p_0(\theta|\bar{e})\big(\boldsymbol{\mu}(\theta, \bar{e}) - \boldsymbol{\mu}(\theta, e)\big) \cdot \boldsymbol{w}(\theta) \geq c(\bar{e}) - c(e)$, $\forall \bar{e} \in E^*$, $\forall e \in E$, and
(d) $w^n(\theta) \geq \underline{w}$, $\forall \theta \in \Theta$, $\forall n = 1, 2, \ldots, N$,

where (a) follows from each type extracting the first-best surplus, (b) follows from P-IC$^{\theta}_{\hat{\theta}}$, (c) follows from A-IC$^{e}_{\bar{e}}$, and (d) follows the limited liability constraints. We get i-iv of Lemma 3 by setting $k^n_\theta = w^n(\theta) - c(e^*(\theta)) - \underline{u}$ for $n = 1, 2, \ldots, N$ and $\theta \in \Theta$.

To see sufficiency, note that if Lemma 3 holds, then there is a mechanism M 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. □

Details on Example 1: Suppose Assumption 1 does not hold. Without loss of generality, let $\theta'' > \theta'$. Let us set $\theta^* = \theta''$. Then, there would exist a vector $\boldsymbol{k} \in \mathbb{R}^N$ solving

$$\underbrace{\begin{bmatrix} \mu_1(\theta^*, e^*(\theta^*)) & \ldots & \mu_N(\theta^*, e^*(\theta^*)) \\ \mu_1(\theta', e^*(\theta^*)) & \ldots & \mu_N(\theta', e^*(\theta^*)) \\ \mu_1(\theta^*, e_1) & \ldots & \mu_N(\theta^*, e_1) \end{bmatrix}}_{A} \begin{bmatrix} k_1 \\ \vdots \\ k_N \end{bmatrix} = \underbrace{\begin{bmatrix} 0 \\ y_1 \\ y_2 \end{bmatrix}}_{C}$$

for some $y_1 \geq S(\theta', e^*(\theta^*)) - S(\theta', e^*(\theta')) = 0$ and $y_2 \leq -c(e^*(\theta^*)) < 0$. However, notice that for $\boldsymbol{\mu}(\theta', e^*(\theta^*)) = \delta\boldsymbol{\mu}(\theta^*, e_1) + (1-\delta)\boldsymbol{\mu}(\theta^*, e^*(\theta^*))$ where $\delta = \frac{\lambda(\theta^*, e^*(\theta^*)) - \lambda(\theta', e^*(\theta^*))}{\lambda(\theta^*, e^*(\theta^*)) - \lambda(\theta^*, e_1)}$. Notice that this implies $rank(A) = 2$. Since we have assumed the system of linear equations has a solution, it also means that the augmented matrix has $rank(A|C) = 2$. However, the latter requires $y_1 = \delta y_2$, which is impossible as $y_1 \geq 0$, $y_2 < 0$ and $\delta > 0$ by the monotonicity assumptions on λ.

References

Beaudry, Paul: Why an informed principal may leave rents to an agent. Int. Econ. Rev. **35**(4), 821–832 (1994)

Grossman, Sanford J., Hart, Oliver D.: An analysis of the principal-agent problem. Econometrica: Journal of the Econometric Society, 7–45 (1983)

Inderst, Roman: Incentive schemes as a signaling device. J. Econ. Behavi. Org. **44**(4), 455–465 (2001)

Karle, Heiko, Schumacher, Heiner, Staat, Christian: Signaling quality with increased incentives. Euro. Econ. Rev. **85**, 8–21 (2016)

Maskin, Eric, Tirole, Jean: The principal-agent relationship with an informed principal: the case of private values. Econometrica **58**(2), 379–409 (1990)

Maskin, Eric, Tirole, Jean: The principal-agent relationship with an informed principal, ii: common values. Econometrica **60**(1), 1–42 (1992)

Myerson, Roger B.: Mechanism design by an informed principal. Econometrica **51**(6), 1767–1797 (1983)

Mylovanov, Tymofiy, Tröger, Thomas: Informed-principal problems in environments with generalized private values. Theoretical Econ. **7**(3), 465–488 (2012)

Mylovanov, Tymofiy, Tröger, Thomas: Mechanism design by an informed principal: private values with transferable utility. Rev. Econ. Stud. **81**(4), 1668 (2014)

Wagner, Christoph, Mylovanov, Tymofiy, Tröger, Thomas: Informed-principal problem with moral hazard, risk neutrality, and no limited liability. J. Econ. Theory **159**, 280–289 (2015)