Contracting with Heterogeneous Beliefs[*]

Xueying Zhao [†]

September 2023

- Click here for the latest version -

Abstract

This paper analyzes the optimal design of incentive contracts in the presence of belief heterogeneity between a principal and an agent. The principal hires the agent to perform a task but cannot observe the agent's actions. Both parties may hold heterogeneous beliefs about the distribution of output realizations. We introduce an "implementation condition" under which the first-best outcomes from the full-information benchmark are attainable, despite information asymmetry. When this condition does not hold, the first-best contract can still be approximated. Additionally, we provide a rationale for the optimality of linear contracts when outputs are normally distributed, and the agent has constant absolute risk aversion preferences.

Keywords: Contract design, heterogeneous beliefs, monotone likelihood ratio property, first-best implementation, linear contracts

JEL Codes: D83, D86, J33

[*] I especially thank Daniel Sgroi and Sinem Hidir for their guidance and support. I also thank Robert Akerlof, Pablo Beker, Dan Bernhardt, Costas Cavounidis, Peter Hammond, Motty Perry, Kirill Pogorelskiy, Herakles Polemarchakis, Philip Reny, and participants at Stony Brook International Conference on Game Theory, Annual Conference of the Israeli Chapter of the Game Theory Society, RES and SES Annual Conference, Lisbon Meetings in Game Theory and Applications, Annual Conference "Contests: Theory and Evidence", and Warwick MIWP Workshop for their helpful comments and suggestions.

[†] Department of Economics, University of Warwick. Email: xueying.zhao@warwick.ac.uk

1 Introduction

The design of optimal contracts is a fundamental concern in organizations, especially when a principal (e.g., employer) hires an agent (e.g., employee) but cannot directly observe the agent's actions. In such situations, performance-based compensation is a commonly used mechanism to encourage the agent to exert effort. While traditional models often assume that the principal and agent share the same beliefs about the returns to effort, in practice, they may hold different beliefs, which can significantly influence the structure of optimal incentive contracts.

This paper explores the optimal design of contracts in the presence of **belief heterogeneity** between a risk-neutral principal and a risk-averse agent. The principal hires an agent to perform a task with the objective of incentivizing the agent to exert high effort, even though the agent's actions are unobservable. While the principal cannot directly monitor the agent's effort, the output of the task is observable, enabling the design of an **output-contingent** payment scheme. Crucially, the principal and agent may hold heterogeneous beliefs regarding the distribution of output realizations. The objective is to design a **cost-minimizing contract** for the principal that **incentivizes high effort** from the agent.

The full-information benchmark, where the principal can observe the agent's actions, serves as a crucial comparison for our analysis. In this scenario, when the principal and agent hold homogeneous beliefs, the principal faces an optimal risk-sharing problem. The principal fully insures the risk-averse agent by bearing all risks, resulting in an optimal contract—referred to as the "**first-best contract**"—that is constant across output realizations. However, when the principal and agent hold heterogeneous beliefs, the principal must balance between risk-sharing and betting on their respective beliefs. Under belief heterogeneity, the first-best contract is no longer constant: the agent receives a higher payment for outputs he believes are more likely than the principal does, and a lower payment for outputs he perceives as less likely. Furthermore, we establish a **monotonicity property** (Proposition 1) of the first-best contract: payment increases with output if the output distribution, as perceived by the agent, dominates that perceived by the principal in the monotone likelihood ratio order. Conversely, payment decreases with output if the output distribution, as perceived by the principal, dominates that perceived by the agent.

In the actual setting, where the principal cannot observe the agent's actions, the principal can rely on the output realizations to infer the agent's chosen effort level. In contrast to the conventional model with homogeneous beliefs, where the first-best contract is unattainable and leads to efficiency loss, belief heterogeneity induces a novel feature. We introduce the "implementation condition" and show that, under this condition, the first-best outcome can be **achieved** (Proposition 2). The intuition is that belief heterogeneity ensures the feasibility of the first-best contract. Even when this condition fails, the first-best contract can still be **approximated**, as extremely low output realizations provide nearly perfect information about the agent's actions (Proposition 4).

Linear contracts, while popular in practice, are generally not optimal in the static setting. However, we provide a rationale for the **optimality of linear contracts** by assuming constant absolute risk aversion preference for the agent and normally distributed outputs, where the distributions differ in means but share the same variance. In this specific case, the first-best contract is linear in outputs (Proposition 5). Moreover, we identify conditions under which the optimal contract coincides with the first-best contract, resulting in a linear payment structure based on output realizations (Proposition 6).

1.1 Related Literature

This paper contributes to the literature on behavioral contract theory (Koszegi, 2014). The existing studies primarily explore the role of overconfident agents in optimal contracting. Santos-Pinto (2008) investigates how agents' mistaken beliefs about their own ability affect the principal's welfare. He shows that a positive self-image held by the agent benefits the principal when effort is observable. However, when effort is unobservable, this positive self-image still favors to the principal, but only under specific conditions. Similarly, De la Rosa (2011) examines the impact of overconfidence on the shape of incentive contracts, identifying two conflicting effects: the incentive effect, where a lower-powered incentive suffices to induce effort, and the wager effect, where a high-powered incentive is preferred by the overconfident agent. The incentive effect dominates when the agent is only slightly overconfident, while the wager effect takes precedence when the agent is significantly overconfident. Santos-Pinto and De la Rosa (2020) provide a formal view of the role of worker overconfidence in internal and external labor markets.

Another relevant literature is exploitative contracting, where the principal has superior information and seeks to exploit the agent's mistakes (DellaVigna and Malmendier, 2004). Fang and Moscarini (2005) examine a principal-agent model with non-common priors, focusing on the signaling function of contracts. In their model, the informed principal knows the true ability of optimistic agents and faces a trade-off between offering appropriate incentives and managing the potential negative impact on productivity when agents learn their true ability. Similarly, Eliaz and Spiegler (2006) explore a principal-agent model with non-common priors but focus on hidden information regarding agents' dynamically inconsistent preferences. Auster (2013) studies how a principal can exploit agents who are unaware of certain possible production outcomes.

This paper also contributes to the literature on linear contracts. Holmstrom and Milgrom (1987) identify conditions under which linear contracts are optimal in a dynamic setting. Carroll (2015) demonstrates that linear contracts can be optimal with risk-neutrality and limited liability, especially when the principal is unaware of the production technology. In contrast to these studies, we derive the optimality of linear contracts from the context of belief heterogeneity.

Additionally, this paper adds to the broader literature on heterogeneous beliefs in economic models. The implications of non-common priors are discussed by Morris (1995) and Hanson (2006). Brunnermeier, Simsek, and Xiong (2014) propose a welfare criterion for models involving heterogeneous beliefs. Alonso and Camara (2016) explore how a sender can design experiments to persuade a receiver when both parties disagree on the likelihood of payoff-relevant states.

2 The Model

Consider two players: a **risk-neutral** principal (P) and a **risk-averse** agent (A). The principal hires the agent to perform a task. The agent can exert effort $e \in \{e_L, e_H\}$, where exerting high effort e_H incurs a cost: $c(e_H) \triangleq c > 0$, and low effort e_L incurs no cost: $c(e_L) \triangleq 0$.

While the principal cannot observe the agent's actions, the output of the task, denoted by $y \in Y \triangleq \mathbb{R} \cup \{-\infty, +\infty\}$, is observable and verifiable. To induce the agent to exert high effort, the principal offers a contract that specifies an output-contingent payment:

$$w(y) : Y \to \mathbb{R},$$

which the agent may accept or reject. If the agent accepts the contract, he chooses his effort level. If the agent rejects the contract, he receives the reservation utility $\underline{u}$ from an outside option. The agent's utility function is additively separable in income and action, given by:

$$u(w(y)) - c(e),$$

where $u(0) = 0$, $u'(\cdot) > 0$, and $u''(\cdot) < 0$. The agent is risk-averse, with diminishing marginal utility from income.

The principal and agent may hold **heterogeneous beliefs** about the distribution of output given effort. The principal believes that the output y is distributed according to $f_P(y \mid e)$, the probability density function for y given effort e, and $F_P(\cdot \mid e)$ is the corresponding cumulative distribution function. Similarly, the agent believes that output follows the distribution $f_A(y \mid e)$, with $F_A(\cdot \mid e)$ as the cumulative distribution function. These distributions are common knowledge and have full support over the set Y.

The principal's objective is to design a **cost-minimizing** contract that incentivizes the agent to exert **high effort** e_H.

Assumption 1 (Monotone Likelihood Ratio Property (MLRP))**.** *Both the principal and agent's probability distributions $f_P(y \mid e)$ and $f_A(y \mid e)$, for $e \in \{e_L, e_H\}$, satisfy the monotone likelihood ratio property:*

$$\frac{\mathrm{d}}{\mathrm{d}y}\left(\frac{f_i(y \mid e_H)}{f_i(y \mid e_L)}\right) \geq 0, \quad \forall y \in Y, \quad i \in \{P, A\}.$$

Under Assumption 1, both the principal and agent believe that higher output realizations imply a higher likelihood that the agent exerted high effort e_H.

Discussion on Heterogeneous Beliefs

A key critique of deviating from the common prior assumption is the perceived lack of discipline in modeling belief heterogeneity. However, model-based approaches help impose meaningful restrictions on the beliefs of different players (Spiegler, 2016; Mailath and Samuelson, 2020). The growing literature on misspecified model offers a rationale for the heterogeneity in beliefs, suggesting that divergent beliefs may arise endogenously from agents fitting subjective causal models to the objective data generating process (Schumacher and Thysen, 2022). In this paper, however, the heterogeneity in prior beliefs is treated as exogenously given.

3 Contract Design

There are no concerns regarding signaling or screening because the probability distributions of output are common knowledge. Although the principal and agent hold different beliefs about the outcome distributions, they "agree to disagree", meaning they base decisions on their own individual beliefs. This distinction is reflected in the subscript of the expectation operator in the principal's cost-minimizing problem, as shown below:

$$\begin{aligned} \underset{w(y)}{\text{minimize}} \quad & \mathbb{E}_P[w(y) \mid e_H] \\ \text{subject to} \quad & \mathbb{E}_A[u(w(y)) \mid e_H] - c \geq \underline{u} \qquad (IR), \\ & \mathbb{E}_A[u(w(y)) \mid e_H] - c \geq \mathbb{E}_A[u(w(y)) \mid e_L] \qquad (IC). \end{aligned} \tag{1}$$

In problem (1), the agent's expected utility from exerting high effort is weakly higher than the reservation utility from an outside option, known as the individual-rationality (IR) constraint. Additionally, the agent's expected utility from exerting high effort is weakly higher than the utility from deviating to low effort, known as the incentive-compatibility (IC) constraint.

3.1 Full-information Benchmark

Consider the full-information benchmark, which serves a useful comparison for our analysis. In this scenario, the principal can observe the agent's actions, eliminating any concerns about incentives. As a result, the principal faces a relaxed optimization problem:

$$\begin{aligned} \underset{w(y)}{\text{minimize}} \quad & \mathbb{E}_P[w(y) \mid e_H] \\ \text{subject to} \quad & \mathbb{E}_A[u(w(y)) \mid e_H] - c \geq \underline{u} \qquad (IR). \end{aligned} \tag{2}$$

Definition 1 (First-best Contract)**.** *The* ***first-best contract****, denoted by* $w^{fb}(y) : Y \to \mathbb{R}$*, is the least costly contract that the principal can offer to incentivize high effort in the full-information benchmark.*

Lemma 1. *The following properties hold in the first-best contract* $w^{fb}(y)$*:*

(i) IR binds:

$$\mathbb{E}_A[u(w^{fb}(y)) \mid e_H] = \underline{u} + c; \tag{3}$$

(ii) The first-order condition is satisfied:

$$\frac{1}{u'(w^{fb}(y))} = \frac{f_A(y \mid e_H)}{f_P(y \mid e_H)}\lambda, \ \ \forall y \in Y, \tag{4}$$

where λ *is the Lagrange multiplier.*

Given additively separable utility function, the agent's IR constraint must bind.[1] Consequently, the agent receives the reservation utility in the first-best contract. The Lagrangian for problem (2) is then given by:

$$\mathcal{L}(\lambda) = \int w(y) f_P(y \mid e_H)\, dy + \lambda \left(\underline{u} + c - \int u(w(y)) f_A(y \mid e_H)\, dy \right),$$

and we can derive the first-order condition.

Homogeneous Beliefs

When the principal and agent share the same beliefs:

$$f_A(y \mid e_H) = f_P(y \mid e_H), \ \ \forall y \in Y,$$

the principal faces an optimal risk-sharing problem. In this case, the first-best contract is constant across all output levels,[2] denoted by $\overline{w}$, where

$$w^{fb}(y) \triangleq \overline{w} = u^{-1}(\underline{u} + c), \ \ \forall y \in Y.$$

This implies that the risk-neutral principal bears full risk, providing full insurance to the risk-averse agent.

Heterogeneous Beliefs

When the principal and agent hold heterogeneous beliefs, the principal faces a trade-off between risk-sharing and betting. Consequently, the first-best contract $w^{fb}(y)$ is no longer constant and exhibits the following properties.

From equation (4) and the fact that $u''(\cdot) < 0$, we can demonstrate that the agent receives a higher payment for outputs that the agent believes are more likely than the principal does, compared to $\overline{w}$:

$$w^{fb}(y) > \overline{w}, \qquad \forall y \in Y \text{ with } f_A(y \mid e_H) > f_P(y \mid e_H).$$

Conversely, the agent receives a lower payment for outputs that the agent believes are less likely than the principal does:

[1] See Proposition 2 in Grossman and Hart (1983).

[2] Assume u is invertible.

$$w^{fb}(y) < \overline{w}, \qquad \forall y \in Y \text{ with } f_A(y \mid e_H) < f_P(y \mid e_H).$$

Additionally, we derive the monotonicity property of the first-best contract as follows.

Proposition 1 (Monotonicity)**.** *The first-best contract is monotonic in outputs if the distributions $f_A(y \mid e_H)$ and $f_P(y \mid e_H)$ satisfy the monotone likelihood ratio property. Specifically, for all $y \in Y$,*

(i) $\frac{\mathrm{d}w^{fb}(y)}{\mathrm{d}y} \geq 0$, *if* $\frac{\mathrm{d}}{\mathrm{d}y}\left(\frac{f_A(y|e_H)}{f_P(y|e_H)}\right) \geq 0$*;*

(ii) $\frac{\mathrm{d}w^{fb}(y)}{\mathrm{d}y} \leq 0$, *if* $\frac{\mathrm{d}}{\mathrm{d}y}\left(\frac{f_A(y|e_H)}{f_P(y|e_H)}\right) \leq 0$.

When the ratio of the agent's perceived probability of output given high effort to the principal's perceived probability is weakly increasing in outputs, higher output realizations imply a higher likelihood of outputs being drawn from the distribution $F_A(\cdot \mid e_H)$. Consequently, the payment increases with outputs. Conversely, when this ratio is weakly decreasing in outputs, higher output realizations imply a lower likelihood of outputs being drawn from the distribution $F_A(\cdot \mid e_H)$, and the payment decreases in outputs. Detailed proofs can be found in the Appendix.

3.2 Optimal Contract

We now discuss the optimal design of contract for problem 1, where the principal cannot observe the agent's actions. In the conventional model where the principal and agent share common prior beliefs, the optimal contracting is a trade-off between risk-sharing and incentives. Once we deviate from the common prior assumption, the optimal choice of incentive scheme makes a trade-off between risk-sharing, incentives and betting.

Definition 2 (Optimal Contract)**.** *The **optimal contract**, denoted by $w^*(y) : Y \to \mathbb{R}$, is the least costly contract that the principal can offer to incentivize high effort when the principal cannot observe the agent's actions.*

Definition 3 (Implementability Condition)**.** *We say that the **implementability condition** holds if*

$$\mathbb{E}_A[u(w^{fb}(y)) \mid e_L] \leq \underline{u}. \tag{5}$$

The equation (5) implies the agent's expected utility from accepting the first-best contract while deviating to low effort is less than the reservation utility.

Proposition 2 (First-best Implementation)**.** *The optimal contract in problem (1) coincides with the first-best contract in problem (2), such that:*

$$w^*(y) = w^{fb}(y), \quad \forall y \in Y.$$

if and only if the implementability condition holds.

The intuition behind this result is that, under the implementation condition, the first-best contract becomes feasible in problem (1). Then, this contract must be optimal in problem (1), since it is the optimal solution for the relaxed problem (2). For detailed proofs, see the Appendix.

Proposition 3. *The implementability condition fails if*

$$\frac{\mathrm{d}}{\mathrm{d}y}\left(\frac{f_A(y \mid e_H)}{f_P(y \mid e_H)}\right) \leq 0, \quad \forall y \in Y. \tag{6}$$

According to Proposition 1, the first-best contract $w^{fb}(y)$ decreases in outputs if the likelihood ratio decreases in outputs. In this case, the agent's expected utility from accepting the contract $w^{fb}(y)$ and subsequently deviating to low effort becomes profitable, thereby violating the implementable condition. Detailed proofs are provided in the Appendix.

From Proposition 3, we derive a necessary condition for the implementability condition to hold: specifically, equation (6) must not hold.

Corollary 1. *If the principal and agent hold the same beliefs, then the first-best contract cannot be attainable in problem (1).*

Proof. If the principal and agent hold the same beliefs:

$$f_A(y \mid e_H) = f_P(y \mid e_H), \forall y \in Y,$$

then:

$$\frac{\mathrm{d}}{\mathrm{d}y}\left(\frac{f_A(y \mid e_H)}{f_P(y \mid e_H)}\right) = 0, \forall y \in Y.$$

Therefore, by Proposition 3, the implementability condition fails. Consequently, by Proposition 2, the first-best contract is not attainable. □

This result illustrates a familiar property of the conventional moral hazard model, where the principal and agent hold homogeneous beliefs. In contrast, with belief heterogeneity, the first-best contract is attainable, as demonstrated in Proposition 2.

Next, we provide conditions under which the first-best contract can be approximated when the implementability condition fails.

Proposition 4 (Approximation)**.** *The first-best contract can be approximated in problem (1), if the following conditions hold:*

$$\frac{\mathrm{d}}{\mathrm{d}y}\left(\frac{f_A(y \mid e_H)}{f_P(y \mid e_H)}\right) \geq 0; \tag{7}$$

and

$$\frac{f_A(y \mid e_H)}{f_A(y \mid e_L)} \to 0 \quad \textit{as} \quad y \to -\infty. \tag{8}$$

The equation (7) ensures that the agent's expected utility from accepting the first-bets contract while deviating to low effort is bounded. The equation (8) implies that, with unbounded support, the outputs at the left tail of the distribution are almost perfectly informative of the action [3]. By setting punishment at extremely low output realizations, which are almost surely avoidable when the agent exerts high effort, the first-best contract can be approximated. Detailed proofs are provided in the Appendix.

4 Optimality of Linear Contracts

Linear contracts are often popular in proactive, yet they are generally not optimal in the static settings. In this section, we provide a rationale for the optimality of linear contracts. To derive closed-form solutions, we consider normally distributed outputs and assume constant absolute risk aversion (CARA) preferences for the agent.

Assumption 2.

(i) The agent exhibits CARA preferences for income:

$$u(w) \triangleq 1 - e^{-rw},$$

where $r \triangleq \frac{-u''(w)}{u'(w)}$ is the CARA coefficient;

(ii) $\underline{u} < 1 - c$;

[3] See (Mirrlees, 1999)

(iii) The principal believes that the outputs under high effort follow a normal distribution with mean μ_P and variance σ^2, denoted by $\mathcal{N}(\mu_P, \sigma^2)$.

(iv) The agent believes that the outputs under high effort are distributed according to $\mathcal{N}(\mu_A, \sigma^2)$;

(v) Both the principal and agent believe that the outputs under low effort are are distributed according to $\mathcal{N}(\mu, \sigma^2)$.

In Assumption 2, condition (i) implies that the agent's expected utility from exerting high effort is strictly less than $1 - c$. Condition (ii) ensures that the agent's reservation utility is bounded from the above, facilitating the possibility of implementing high effort. Conditions (iii), (iv), and (v) ensure that the normal distributions of output have different means but same variance. With the normal distributions specified, Assumption 1 now becomes $\mu_A \geq \mu$ and $\mu_P \geq \mu$.

We now derive a closed-form solution for the first-best contract.

Proposition 5. *Under Assumption 2, the first-best contract $w^{fb}(y)$ is given by*

$$w^{fb}(y) = \frac{\mu_A - \mu_P}{r\sigma^2} y + \frac{1}{r}\left(\frac{\mu_P^2 - \mu_A^2}{2\sigma^2} + \ln\left(\frac{1}{1 - \underline{u} - c}\right)\right), \forall y \in Y.$$

From Proposition 5, the first-best contract is **linear** in outputs,[4] as the marginal change in payment with respect to output is constant:

$$\frac{\mathrm{d}w^{fb}(y)}{\mathrm{d}y} = \frac{\mu_A - \mu_P}{r\sigma^2}.$$

Specifically, the first-best contract increases linearly in outputs if $\mu_A > \mu_P$ and decreases linearly in outputs if $\mu_A < \mu_P$. Furthermore, the first-best contract becomes more dispersed with greater belief heterogeneity (when $|\mu_A - \mu_P|$ increases), lower risk aversion of the agent (r decreases) or less noise in the output distribution (when σ decreases).

Proposition 6 (Optimality of Linear Contract)**.** *Under Assumption 2, the optimal contract is linear in outputs if the following conditions hold:*

$$\frac{(\mu_A - \mu_P)(\mu_A - \mu)}{\sigma^2} \geq \ln\left(\frac{1 - \underline{u}}{1 - \underline{u} - c}\right), \tag{9}$$

and

$$\frac{(\mu_A - \mu_P)^2}{2\sigma^2} + r(\mu_P - \mu) \geq \ln\left(\frac{1 - \underline{u}}{1 - \underline{u} - c}\right). \tag{10}$$

[4] This result can be generalized into the natural exponential family, represented as $f(y \mid \theta) = h(y)\exp(\theta y - A(\theta))$, which includes normal distributions with known variance.

In Proposition 6, we establish conditions under which a linear contract is optimal. This conclusion arises from two key factors: first, the first-best contract is linear in outputs under CARA preferences and normally distributed outputs, as outlined in Assumption 2; second, belief heterogeneity enhances the feasibility of the first-best contract. Condition (9) ensures that the implementation condition holds. Then, by Proposition 2, the optimal contract coincides with the first-best contract. Condition (10) guarantees that the profit for the principal from using the first-best contract to incentivize high effort is weakly higher the optimal profit from incentivizing low effort. With greater belief heterogeneity (when $|\mu_A - \mu_P|$ increases), it becomes easier to incentivize the agent to exert high effort, thus increasing the likelihood of a linear optimal contract. A graphical illustration follows.

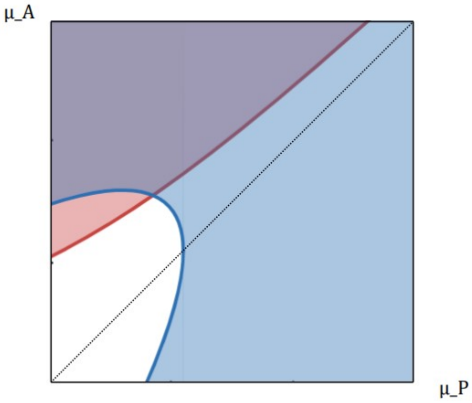

Figure 1: Optimality of Linear Contract.

In Figure 1, we set the parameters as follows: $\mu = 0, \sigma^2 = 1, \underline{u} = 0, r = 1, c = 2/3$. In this example, condition (9) simplifies to

$$\mu_A(\mu_A - \mu_P) \geq \ln 3.$$

This means that when the pair (μ_p, μ_A) falls within the red region of the graph, the optimal contract required to implement high effort coincides with the first-best contract, which is linear in outputs. For condition (10), we have

$$\frac{(\mu_A - \mu_P)^2}{2} + \mu_P \geq \ln 3.$$

Hence, when (μ_p, μ_A) is located in the blue region, the principal prefers to implement high effort via the first-best contract over low effort. Consequently, the optimal contact is linear in outputs if the pair (μ_p, μ_A) falls within the purple region, which represents the intersection of red and blue regions.

5 Conclusion

In this paper, we analyze the design of optimal contracts aimed at incentivizing high effort in a model where the principal and agent hold heterogeneous beliefs about the distribution of output realizations. We introduce the implementation condition and demonstrate that, under this condition, the first-best outcomes from the full-information benchmark can be achieved, despite the information asymmetry regarding the agent's actions. Furthermore, we provide a rationale for the optimality of linear contracts in the context of normally distributed outputs and constant absolute risk aversion preferences.

Appendix

PROOF OF PROPOSITION 1.

Applying the implicit function theorem to (4), we get

$$\frac{\mathrm{d}w^{fb}(y)}{\mathrm{d}y} = \frac{\frac{\mathrm{d}}{\mathrm{d}y}\left(\ln\left(\frac{f_A(y|e_H)}{f_P(y|e_H)}\right)\right)}{\frac{-u''(w^{fb}(y))}{u'(w^{fb}(y))}}. \tag{A1}$$

Given that $u'(\cdot) > 0$ and $u''(\cdot) < 0$, if the likelihood ratio, $\frac{f_A(y|e_H)}{f_P(y|e_H)}$, is weakly increasing in outputs, then the payment increases with outputs due to the non-negativity of the expression on the right-hand side of (A1). Conversely, if the likelihood ratio is weakly decreasing in outputs, then the payment decreases with outputs. □

PROOF OF PROPOSITION 2.

We first prove the feasibility of the first-best contract $w^{fb}(y)$ in problem (1). From (4), we have

$$\mathbb{E}_A[u(w^{fb}(y)) \mid e_H] - c = \underline{u}, \tag{A2}$$

which implies that the contract $w^{fb}(y)$ is individual-rational.

Under the implementability condition, we have (5), which combined with (A2), leads to

$$\mathbb{E}_A[u(w^{fb}(y)) \mid e_H] - c \geq \mathbb{E}_A[u(w^{fb}(y)) \mid e_L],$$

implying that the contract $w^{fb}(y)$ is incentive-compatible.

Therefore, the first-best contract is feasible for problem (1). Furthermore, the first-best contract must be optimal in problem (1), since it is the optimal solution for the relaxed problem (2). The other direction is straightforward to verify. □

Lemma 2. *The monotone likelihood ratio property implies the first-order stochastic dominance: that is, for all $y \in Y$,*

$$\frac{\mathrm{d}}{\mathrm{d}y}\left(\frac{f_A(y \mid e_H)}{f_A(y \mid e_L)}\right) \geq 0 \;\Rightarrow\; F_A(y \mid e_L) \geq F_A(y \mid e_H).$$

PROOF OF LEMMA 2.

If $\frac{\mathrm{d}}{\mathrm{d}y}\left(\frac{f_A(y|e_H)}{f_A(y|e_L)}\right) \geq 0$ for all y, then for $y_1 > y_2$, we have $\frac{f_A(y_1|e_H)}{f_A(y_1|e_L)} \geq \frac{f_A(y_2|e_H)}{f_A(y_2|e_L)}$, which can be rewritten as:

$$f_A(y_1 \mid e_H) f_A(y_2 \mid e_L) \geq f_A(y_2 \mid e_H) f_A(y_1 \mid e_L). \tag{A3}$$

Integrating y_2 from $-\infty$ to y_1 on both sides of equation (A3), we have:

$$f_A(y_1 \mid e_H) F_A(y_1 \mid e_L) \geq F_A(y_1 \mid e_H) f_A(y_1 \mid e_L),$$

which can be rewritten as:

$$\frac{f_A(y_1 \mid e_H)}{f_A(y_1 \mid e_L)} \geq \frac{F_A(y_1 \mid e_H)}{F_A(y_1 \mid e_L)}.$$

Integrating y_1 from y_2 to ∞ on both sides of equation (A3), we have:

$$[1 - F_A(y_2 \mid e_H)] f_A(y_2 \mid e_L) \geq f_A(y_2 \mid e_H)[1 - F_A(y_2 \mid e_L)],$$

which can be rewritten as:

$$\frac{1 - F_A(y_2 \mid e_H)}{1 - F_A(y_2 \mid e_L)} \geq \frac{f_A(y_2 \mid e_H)}{f_A(y_2 \mid e_L)}.$$

Therefore, for any arbitrary y, we have:

$$\frac{1 - F_A(y \mid e_H)}{1 - F_A(y \mid e_L)} \geq \frac{f_A(y \mid e_H)}{f_A(y \mid e_L)} \geq \frac{F_A(y \mid e_H)}{F_A(y \mid e_L)},$$

which implies that:

$$F_A(y \mid e_L) \geq F_A(y \mid e_H).$$

□

PROOF OF PROPOSITION 3.

Integrating by parts, we find:

$$\begin{aligned}\mathbb{E}_A[u(w(y)) \mid e] &= \int u(w(y))\, dF_A(y \mid e) \\ &= u(w(y))F_A(y \mid e) - \int u'(w(y))w'(y)F_A(y \mid e)\, dy.\end{aligned}$$

For any **arbitrary** contract $w(y)$, the difference between the agent's expected utility from high effort and low effort is given by

$$\begin{aligned}
&\mathbb{E}_A[u(w(y)) \mid e_H] - \mathbb{E}_A[u(w(y)) \mid e_L] \\
&= -\int u'(w(y))w'(y)F_A(y \mid e_H)\,dy + \int u'(w(y))w'(y)F_A(y \mid e_L)\,dy \\
&= \int u'(w(y))w'(y)[F_A(y \mid e_L) - F_A(y \mid e_H)]\,dy. \qquad \text{(A4)}
\end{aligned}$$

Assume that

$$\frac{\mathrm{d}}{\mathrm{d}y}\left(\frac{f_A(y \mid e_H)}{f_P(y \mid e_H)}\right) \leq 0,\ \forall y \in Y.$$

Then, from (*ii*) of Proposition 1, the first-best contract decreases in outputs:

$$\frac{\mathrm{d}w^{fb}(y)}{\mathrm{d}y} \leq 0,\ \forall y \in Y. \qquad \text{(A5)}$$

Under Assumption 1, $\frac{\mathrm{d}}{\mathrm{d}y}\left(\frac{f_A(y|e_H)}{f_A(y|e_L)}\right) \geq 0, \forall y \in Y$. Then, by Lemma 2, we have

$$F_A(y \mid e_L) \geq F_A(y \mid e_H),\ \forall y \in Y. \qquad \text{(A6)}$$

From (A4), the difference in agent's expected utility from the first-best contract is

$$\mathbb{E}_A[u(w^{fb}(y)) \mid e_H] - \mathbb{E}_A[u(w^{fb}(y)) \mid e_L] = \int u'(w^{fb}(y))\frac{\mathrm{d}w^{fb}(y)}{\mathrm{d}y}[F_A(y \mid e_L) - F_A(y \mid e_H)]\,dy.$$

Then, from $u'(w) > 0$, (A5) and (A6), we have

$$\mathbb{E}_A[u(w^{fb}(y)) \mid e_H] - \mathbb{E}_A[u(w^{fb}(y)) \mid e_L] \leq 0,$$

which leads to:

$$\mathbb{E}_A[u(w^{fb}(y)) \mid e_L] \geq \mathbb{E}_A[u(w^{fb}(y)) \mid e_H] = \underline{u} + c,$$

where the equality comes from (3).

Since $c > 0$, we have:

$$\mathbb{E}_A[u(w^{fb}(y)) \mid e_L] > \underline{u},$$

which implies that the implementability condition fails. □

PROOF OF PROPOSITION 4.

By definition, when the implementability condition fails, the first-best contract $w^{fb}(y)$ must satisfy the following equation:

$$\int_{-\infty}^{\infty} u(w^{fb}(y)) f_A(y \mid e_L)\, dy > \underline{u}. \tag{A7}$$

Then, rewrite equation (3) as:

$$\int_{-\infty}^{\infty} u(w^{fb}(y)) f_A(y \mid e_H)\, dy = \underline{u} + c, \tag{A8}$$

and we get

$$\int_{-\infty}^{\infty} u(w^{fb}(y))(f_A(y \mid e_H) - f_A(y \mid e_L))\, dy < c,$$

which implies that the first-best contract violates the IC constraint.

Consider an alternative contract $\tilde{w}(y)$:

$$\tilde{w}(y) = \begin{cases} w^{fb}(y), & \text{if } y > \underline{y}; \\ K, & \text{if } \underline{\underline{y}} \leq y \leq \underline{y}; \\ w^{fb}(y), & \text{if } y < \underline{\underline{y}}. \end{cases}$$

In this contract $\tilde{w}(y)$, the principal will punish the agent (with K set sufficiently low), if the output realizations are extremely low and fall into the interval $[\underline{\underline{y}}, \underline{y}]$. For all other output realizations, the agent receives the same payment as in the first-best contract.

For any arbitrary outputs $\underline{\underline{y}}$ and $\underline{y}$, where $f_A(y \mid e_H) < f_A(y \mid e_L)$ for $y \in [\underline{\underline{y}}, \underline{y}]$, we can construct a payment K, where $u(K) < u(w^{fb}(y))$ for $y \in [\underline{\underline{y}}, \underline{y}]$, such that the following equation holds:

$$\begin{aligned} &\int_{-\infty}^{\underline{\underline{y}}} u(w^{fb}(y))(f_A(y \mid e_H) - f_A(y \mid e_L))\, dy + \int_{\underline{\underline{y}}}^{\underline{y}} u(K)(f_A(y \mid e_H) - f_A(y \mid e_L))\, dy \\ &+ \int_{\underline{y}}^{\infty} u(w^{fb}(y))(f_A(y \mid e_H) - f_A(y \mid e_L))\, dy = c, \end{aligned} \tag{A9}$$

which ensures that the contract $\tilde{w}(y)$ satisfies the IC constraint.

Next we check the IR constraint for the contract $\tilde{w}(y)$.

The difference in agent's expected utility between the outside option and the contract $\tilde{w}(y)$ is given by

$$\begin{aligned}
&\underline{u} - \left(\int_{-\infty}^{\infty} u(\tilde{w}(y)) f_A(y \mid e_H)\, dy - c\right) \\
&= \left(\int_{-\infty}^{\infty} u(w^{fb}(y)) f_A(y \mid e_H)\, dy - c\right) - \left(\int_{-\infty}^{\infty} u(\tilde{w}(y)) f_A(y \mid e_H)\, dy - c\right) \\
&= \int_{\underline{\underline{y}}}^{\underline{y}} (u(w^{fb}(y)) - u(K)) f_A(y \mid e_H)\, dy, \qquad \text{(A10)}
\end{aligned}$$

where the first equality comes from (A8).

Under Assumption 1, we have $\frac{\mathrm{d}}{\mathrm{d}y}\left(\frac{f_A(y|e_H)}{f_A(y|e_L)}\right) \geq 0$.

If $\frac{f_A(y|e_H)}{f_A(y|e_L)} \to 0$ as $y \to -\infty$, then for any arbitrarily small m (where $0 < m < 1$), there exists an output $\underline{y}$ that is sufficiently low such that $\frac{f_A(y|e_H)}{f_A(y|e_L)} < m$ for all $y < \underline{y}$. This implies

$$f_A(y \mid e_H) < \frac{m}{m-1}(f_A(y \mid e_H) - f_A(y \mid e_L)).$$

Therefore, the difference in agent's expected utility, as in formulation (A10), is strictly less than

$$\frac{m}{m-1}\int_{\underline{\underline{y}}}^{\underline{y}} (u(w^{fb}(y)) - u(K))(f_A(y \mid e_H) - f_A(y \mid e_L))\, dy,$$

which, using equation (A9), is equal to

$$\frac{m}{m-1}\left(\int_{-\infty}^{\infty} u(w^{fb}(y))(f_A(y \mid e_H) - f_A(y \mid e_L))\, dy - c\right),$$

which, using (A8), can be reduced into

$$\frac{m}{m-1}\left(\underline{u} - \int_{-\infty}^{\infty} u(w^{fb}(y)) f_A(y \mid e_L)\, dy\right). \qquad \text{(A11)}$$

The above formulation is strictly positive, from equation (A7) and $\frac{m}{m-1} < 0$.

If $\frac{\mathrm{d}}{\mathrm{d}y}\left(\frac{f_A(y|e_H)}{f_P(y|e_H)}\right) \geq 0,\ \forall y \in Y$, then from (i) in Proposition 1, we have

$$\frac{\mathrm{d}w^{fb}(y)}{\mathrm{d}y} \geq 0.$$

Then, $u(w^{fb}(y))$ weakly increases in y since $u'(\cdot) > 0$.

By Lemma 2, $\frac{\mathrm{d}}{\mathrm{d}y}\left(\frac{f_A(y|e_H)}{f_P(y|e_H)}\right) \geq 0,\ \forall y \in Y$ implies that $F_A(y \mid e_H)$ first-order stochastically dominates $F_A(y \mid e_L)$. Therefore, we have

$$\int_{-\infty}^{\infty} u(w^{fb}(y))f_A(y \mid e_L)\,dy \leq \int_{-\infty}^{\infty} u(w^{fb}(y))f_A(y \mid e_H)\,dy,$$

which, using (A8), is equal to

$$\int_{-\infty}^{\infty} u(w^{fb}(y))f_A(y \mid e_L)\,dy \leq \underline{u} + c.$$

Thus, the difference in the agent's utility between the outside option and the first-best contract is bounded from below:

$$\underline{u} - \int_{-\infty}^{\infty} u(w^{fb}(y))f_A(y \mid e_L)\,dy \geq -c.$$

As $m \to 0$, we have

$$\frac{m}{m-1}\left(\underline{u} - \int_{-\infty}^{\infty} u(w^{fb}(y))f_A(y \mid e_L)\,dy\right) \to 0.$$

This implies that as m is small enough, the difference between agent's expected utility from outside option and $\tilde{w}(y)$ tends to be zero, and thus the first-best contract can be approximated (but never achieved). □

PROOF OF PROPOSITION 5.

From (*iii*) and (*iv*) in Assumption 2, the ratio of $f_A(y \mid e_H)$ to $f_P(y \mid e_H)$ is given by:

$$\frac{f_A(y \mid e_H)}{f_P(y \mid e_H)} = \exp\left(\frac{\mu_A - \mu_P}{\sigma^2}y + \frac{\mu_P^2 - \mu_A^2}{2\sigma^2}\right). \tag{A12}$$

From (*i*) in Assumption 2, $u(w) = 1 - e^{-rw}$, we have:

$$u'(w) = re^{-rw}, \tag{A13}$$

which implies that

$$e^{-rw} = \frac{1}{r}u'(w).$$

Then, rewrite $u(w)$ as:

$$u(w) = 1 - \frac{1}{r}u'(w). \tag{A14}$$

Substituting (A14) into (3), we get:

$$\int \left(1 - \frac{1}{r}u'(w^{fb}(y))\right) dF_A(y \mid e_H) = \underline{u} + c.$$

By (4), we can replace $u'(w^{fb}(y))$ with $\frac{1}{\lambda}\frac{f_P(y|e_H)}{f_A(y|e_H)}$ in the above equation and get:

$$\int \left(1 - \frac{1}{r}\frac{1}{\lambda}\frac{f_P(y \mid e_H)}{f_A(y \mid e_H)}\right) dF_A(y \mid e_H) = \underline{u} + c,$$

which can be rewritten as:

$$\int f_A(y \mid e_H)dy - \frac{1}{r\lambda}\int \frac{f_P(y \mid e_H)}{f_A(y \mid e_H)} f_A(y \mid e_H)dy = \underline{u} + c,$$

which can be simplified into

$$\lambda = \frac{1}{r(1 - \underline{u} - c)}. \tag{A15}$$

From (ii) in assumption 2, we know λ is strictly positive.

Then, substituting (A12) and (A13) into (4), we have:

$$\frac{1}{r}\exp\Big(rw^{fb}(y)\Big) = \lambda \exp\left(\frac{\mu_A - \mu_P}{\sigma^2}y + \frac{\mu_P^2 - \mu_A^2}{2\sigma^2}\right),$$

which can be rewritten as:

$$w^{fb}(y) = \frac{1}{r}\left(\frac{\mu_A - \mu_P}{\sigma^2}y + \frac{\mu_P^2 - \mu_A^2}{2\sigma^2} + \ln(r\lambda)\right).$$

Substituting equation (A15) into the above expression, the first-best contact is given by:

$$w^{fb}(y) = \frac{\mu_A - \mu_P}{r\sigma^2}y + \frac{1}{r}\left(\frac{\mu_P^2 - \mu_A^2}{2\sigma^2} + \ln\left(\frac{1}{1 - \underline{u} - c}\right)\right).$$

□

PROOF OF PROPOSITION 6.

We first calculate the agent's expected utility from accepting the the first-best contract while deviating to low effort, which is given by:

$$\mathbb{E}_A[u(w^{fb}(y)) \mid e_L] = \mathbb{E}_A[1 - e^{-rw^{fb}(y)} \mid e_L], \tag{A16}$$

where the equality comes from equation (A14).

Consider a random variable $z \triangleq e^{-rw^{fb}(y)}$.

From the first-best contract solved in Proposition 5, we get:

$$\ln(z) = -rw^{fb}(y) = -\frac{\mu_A - \mu_P}{\sigma^2}y - \left(\frac{\mu_P^2 - \mu_A^2}{2\sigma^2} + \ln\left(\frac{1}{1-\underline{u}-c}\right)\right).$$

Under Assumption 2, the agent believes that outputs y under low effort are distributed according to the normal distribution $\mathcal{N}(\mu, \sigma^2)$. From the agent's perspective, the random variable $\ln(z)$, which is a linear transformation of y, is thus distributed according to a normal distribution with mean μ_0 and σ_0^2, where

$$\mu_0 \triangleq \mathbb{E}_A[\ln(z) \mid e_L] = \frac{(\mu_A - \mu_P)(\mu_A + \mu_P - 2\mu)}{2\sigma^2} + \ln(1-\underline{u}-c), \tag{A17}$$

$$\sigma_0^2 \triangleq \mathbb{E}_A[(\ln(z) - \mu_0)^2 \mid e_L] = \frac{(\mu_A - \mu_P)^2}{\sigma^2}. \tag{A18}$$

From the agent's perspective, z follows a log-normal distribution, where the mean of z is given by:

$$\mathbb{E}_A[z \mid e_L] \triangleq \exp\big(\mu_0 + \sigma_0^2/2\big).$$

From (A17) and (A18), we get

$$\mu_0 + \sigma_0^2/2 = \frac{(\mu_A - \mu_P)(\mu_A - \mu)}{\sigma^2} + \ln(1-\underline{u}-c),$$

implying

$$\mathbb{E}_A[z \mid e_L] = (1-\underline{u}-c)\exp\left(\frac{(\mu_A - \mu_P)(\mu_A - \mu)}{\sigma^2}\right).$$

From (A16), we have

$$\mathbb{E}_A[u(w^{fb}(y)) \mid e_L] = 1 - \mathbb{E}_A[z \mid e_L] = 1 - (1-\underline{u}-c)\exp\left(\frac{(\mu_A - \mu_P)(\mu_A - \mu)}{\sigma^2}\right).$$

The implementability condition holds if

$$\mathbb{E}_A[u(w^{fb}(y)) \mid e_L] \le \underline{u},$$

which can be simplified into

$$\frac{(\mu_A - \mu_P)(\mu_A - \mu)}{\sigma^2} \ge \ln\left(\frac{1-\underline{u}}{1-\underline{u}-c}\right). \tag{A19}$$

Then, according to Proposition 2, the optimal contract coincides with the first-best contract and thus linear in outputs.

Next, we will provide condition to ensure that the principal prefers to implement high effort rather than low effort.

The principal believes that outputs y under high effort are distributed according to $\mathcal{N}(\mu_P, \sigma^2)$. Thus, he perceives that the expected cost of the first-best contract to implement e_H is given by

$$\mathbb{E}_P[w^{fb}(y) \mid e_H] = \frac{-(\mu_A - \mu_P)^2}{2r\sigma^2} + \frac{1}{r}\ln\left(\frac{1}{1-\underline{u}-c}\right).$$

Note that the cost-minimizing contract to implement e_L is constant across outputs: $\tilde{w} = \frac{1}{r}\ln\left(\frac{1}{1-\underline{u}}\right)$, and the principal weakly prefers to implementing high effort via the first-best contract rather than low effort, if his expected profit from high effort is higher than the profit from low effort:

$$\mathbb{E}_P[y - w^{fb}(y) \mid e_H] \geq \mathbb{E}_P[y - \tilde{w} \mid e_L],$$

which can be reduced into:

$$\mu_P - \mathbb{E}_P[w^{fb}(y) \mid e_H] \geq \mu - \frac{1}{r}\ln\left(\frac{1}{1-\underline{u}}\right),$$

which can be rewritten as:

$$\frac{(\mu_A - \mu_P)^2}{2\sigma^2} + r(\mu_P - \mu) \geq \ln\left(\frac{1-\underline{u}}{1-\underline{u}-c}\right). \tag{A20}$$

Therefore, if equations (A19) and (A20) hold, the optimal contract is linear in outputs. □

References

Alonso, R. and O. Camara (2016). Bayesian persuasion with heterogeneous priors. *Journal of Economic Theory 165*, 672–706.

Auster, S. (2013). Asymmetric awareness and moral hazard. *Games and Economic behavior 82*, 503–521.

Brunnermeier, M. K., A. Simsek, and W. Xiong (2014). A welfare criterion for models with distorted beliefs. *The Quarterly Journal of Economics 129*(4), 1753–1797.

Carroll, G. (2015). Robustness and linear contracts. *American Economic Review 105*(2), 536–563.

De la Rosa, L. E. (2011). Overconfidence and moral hazard. *Games and Economic Behavior 73*(2), 429–451.

DellaVigna, S. and U. Malmendier (2004). Contract design and self-control: Theory and evidence. *The Quarterly Journal of Economics 119*(2), 353–402.

Eliaz, K. and R. Spiegler (2006). Contracting with diversely naive agents. *The Review of Economic Studies 73*(3), 689–714.

Fang, H. and G. Moscarini (2005). Morale hazard. *Journal of Monetary Economics 52*(4), 749–777.

Grossman, S. J. and O. D. Hart (1983). An analysis of the principal-agent problem. *Econometrica 51*, 7–45.

Hanson, R. (2006). Uncommon priors require origin disputes. *Theory and Decision 61*(4), 319–328.

Holmstrom, B. and P. Milgrom (1987). Aggregation and linearity in the provision of intertemporal incentives. *Econometrica: Journal of the Econometric Society*, 303–328.

Koszegi, B. (2014). Behavioral contract theory. *Journal of Economic Literature 52*(4), 1075–1118.

Mailath, G. J. and L. Samuelson (2020). Learning under diverse world views: Model-based inference. *American Economic Review 110*(5), 1464–1501.

Mirrlees, J. A. (1999). The theory of moral hazard and unobservable behaviour: Part i. *The Review of Economic Studies 66*(1), 3–21.

Morris, S. (1995). The common prior assumption in economic theory. *Economics and Philosophy 11*, 227–253.

Santos-Pinto, L. (2008). Positive self-image and incentives in organisations. *The Economic Journal 118*(531), 1315–1332.

Santos-Pinto, L. and L. E. De la Rosa (2020). Overconfidence in labor markets. In K. Zimmermann (Ed.), *Handbook of Labor, Human Resources and Population Economics*. Springer-Verlag.

Schumacher, H. and H. C. Thysen (2022). Equilibrium contracts and boundedly rational expectations. *Theoretical Economics 17*(1), 371–414.

Spiegler, R. (2016). Bayesian networks and boundedly rational expectations. *The Quarterly Journal of Economics 131*(3), 1243–1290.